Retracing the evolution of enzyme regulation

A thesis submitted in partial fulfilment of the requirements

for the degree of Doctor of Philosophy in Biochemistry

at the University of Canterbury

by Fiona Given

2018

Abstract

α-Isopropylmalate synthase (IPMS) catalyses the first committed step in the leucine biosynthesis pathway in microorganisms and some plants. It catalyses the condensation of ketoisovalerate (KIV) and acetyl-coenzyme A (AcCoA) to form isopropylmalate and coenzyme A (CoA). IPMS is allosterically inhibited by the product of the pathway, L-leucine. Structurally, IPMS is a homodimer, and each chain consists of a N-terminal $(\alpha/\beta)_8$ barrel where the active site is located, a catalytic accessory unit formed of subdomain I and subdomain II, and a C-terminal regulatory domain that binds L-leucine. Truncation of IPMS that removes subdomain II or part of subdomain II abolishes catalysis.[1, 2]

Of particular interest in this thesis is IPMS from *Neisseria meningitidis* (*Nme*IPMS). Although this enzyme has been extensively studied, there is no full-length crystal structure available, although there are several of a related enzyme, IPMS from *Mycobacterium tuberculosis* (*Mtu*IPMS), with KIV, and with L-leucine, bound. There is no substantial conformational change observed when the substrate-bound crystal structure is compared to the structure with L-leucine bound. Untangling the dynamic nature of allostery, and how it has evolved, in these proteins is a particular focus of this thesis.

There are also structurally similar proteins that catalyse similar reactions that are also of interest. Citramalate synthase (CMS) is structurally similar to IPMS but catalyses the reaction of pyruvate and AcCoA to form citramalate and CoA in an isoleucine biosynthesis pathway and is inhibited by L-isoleucine. Homocitrate synthase (HCS) functions in a lysine biosynthesis pathway in some organisms and utilises ketoglutarate and AcCoA to form homocitrate and CoA. HCS contains a homologous catalytic domain and catalytic accessory unit as IPMS but lacks a regulatory domain and is competitively inhibited by lysine. The similarities, differences, and modularity of these proteins is explored using computational methods and also by the construction of truncated and fusion proteins.

Chapter 2 utilises a computational method, statistical coupling analysis, to identify a potential network in *Nme*IPMS-like IPMS proteins. Subsequent alanine mutations in *Nme*IPMS demonstrated that mutating charged residues in this proposed network can abolish or attenuate the allosteric signal, suggesting that the network identified may represent a way the allosteric signal is transferred from the allosteric site to the active site. Isothermal titration calorimetry is also used to explore the thermodynamics of L-leucine binding to the wild-type *Nme*IPMS and to the L-leucine insensitive alanine mutants.

Chapter 3 broadens the scope of statistical coupling analysis (SCA) and also uses another computational method, mutual information (MI), to investigate how structurally similar subdomains facilitate catalysis in the presence and absence of a regulatory domain. A population of proteins that contain a regulatory domain, and a population that do not, were assessed using SCA and MI to determine whether there were differences, particularly in the subdomains, that may provide information about maintenance of the balance of flexibility and stability that is crucial to catalysis in these proteins.

Chapter 4 uses an active, truncated, form of *Nme*IPMS to compare and contrast with the wild-type protein. The kinetics of both the truncated *Nme*IPMS and the wild-type *Nme*IPMS are investigated under crowded conditions to explore the impact that viscosity has on these dynamic proteins. Alanine mutations are also made in subdomains I and II to investigate the role of particular residues in catalysis and allostery, and these allow comparison with previous work done on *Mtu*IPMS that highlights the difference between two structurally similar groups of IPMS proteins.

Chapter 5 describes the cloning, expression, and partial purification of an HCS, *Sso*HCS, from *Sulfolobus solfataricus* that appears to have a different type of regulatory domain to the canonical IPMS/CMS regulatory domain. The partial characterisation of this protein suggests that an allosterically regulated HCS has been identified. Chapter 5 also describes the construction of several fusion proteins, where parts of IPMS, HCS, and CMS, are fused to together to explore the modularity of these proteins. Catalysis was preserved in some of the fusions although allostery was not preserved in any so far investigated.

The final chapter includes a broad summary of the work in this thesis as well as ideas for future research. This chapter also contains a discussion about the considerable differences between some IPMS enzymes that, although they catalyse the same reaction, are considerably different taxonomically. Additionally, the important role of networks of residues that facilitate catalysis and allostery is analysed.

Acknowledgements

Firstly, to Professor Emily Parker. Thank you for all your hard work, robust discussion, and allowing me to explore what I found interesting and thus providing me with so many opportunities to learn.

To Team IPMS, past and present, particularly Andrew, Matt, and Wanting. Your enthusiasm (or valid lack thereof at times!) and support while dealing with these wily enzymes was invaluable. Plus helping me fix the BioRad. A lot.

To the Parker Group, especially Emma, Effie, Gerd, Nicky, Kyle, and Leyla. Thank you for your help, your support, and all the encouragement you have given me. Thank you to Annette for helping so much with cloning in the early days.

Thank you to the technical department of the Department of Chemistry/School of Chemical and Physical Sciences for their expertise in fixing all and sundry.

Thank you to the University of Canterbury for the Doctoral Scholarship that enabled me to do this work and to the Biomolecular Interaction Centre and the Marsden Fund for additional funding.

To my friends, my family, and my horses. Thank you for putting up with me over the past few years and keeping me sane!

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AAA	α-aminoadipate
AcCoA	Acetyl-CoenzymeA
CDF	Cumulative density function
CLANS	CLuster ANalysis of Sequences
CMS	Citramalate synthase
DAP	L,L-diaminopimelate
DTP	4'-4'-dithiopyridine
FRET	Fluorescence resonance energy transfer
HCS	Homocitrate synthase
HDX	Hydrogen-deuterium exchange
IPMS	Isopropylmalate synthase
KG	α-ketoglutarate
KIV	α-ketoisovalerate
Lbi	*Leptospira biflexa*
Lin	*Leptospira interrogans*
Nme	*Neisseria meningitidis*
NMR	Nuclear magnetic resonance
MD	Molecular dynamics
MI/MIp	Mutual information
MSA	Multiple sequence alignment
Mtu	*Mycobacterium tuberculosis*
PDB	Protein Data Bank
RMSD	Root mean squared deviation
RDA	Regulatory domain absent
RDP	Regulatory domain present
SAXS	Small-angle X-ray scattering
SCA	Statistical coupling analysis
Spo	*Schizosaccharyomyces pombe*
Sso	*Sulfolobus solfataricus*
TIM	Triosephosphate isomerase
Tth	*Thermus thermophilus*

Chapter 1: Introduction

1.1 Protein dynamics and their importance

Proteins were once thought to be static entities, an idea propagated by the fixed models produced by X-ray crystallography that did not represent the dynamic nature of proteins in the cellular context. It is now apparent that proteins are inherently conformationally flexible through a combination of thermodynamic movement and coordinated motion that can be crucial for processes such as catalysis and allostery.[3-6] Intrinsic protein dynamics can also be important for protein-protein interaction.[7]

1.1.1 Types of protein motion

The time scale upon which proteins move can vary depending on the nature of the motion (Figure 1.1). At the fastest extreme of currently investigated motion is local flexibility, which is typically on the femto-second to pico-second timescale.[8] This type of motion can be revealed via the B-factors in high-resolution X-ray crystallography and by nuclear magnetic resonance (NMR).[9] Local flexibility can be described as highly localised, low energy, movement of the backbone atoms, particularly the C_α-C_β bond, to allow for plasticity in side-chain movement. Moving further down the time scale includes loop motions that can occur over a longer timescale than local flexibility. The movement of flexible loops can be crucial for catalysis, as observed in dihydrofolate reductase from *Escherichia coli* where the movement of the loop region plays a key role in switching between the five kinetic intermediate states during catalysis.[10] Loop regions can have additional roles, such as in the M_2 muscarinic acetylcholine receptor, a G-protein coupled receptor, where a flexible loop acts as a gatekeeper for both the allosteric and orthosteric ligands.[11]

Multi-domain proteins can also undergo domain motions where domains, typically attached by flexible loops or linkers, can move relative to each other. These motions can be important for allosteric regulation. One example is found in the 3-deoxy-*D*-arabino-heptulosonate 7-phosphate synthase from *Thermotoga maritima,* where the regulatory domains shift upon binding of the allosteric inhibitor, tyrosine, to physically block access to the active site.[12] The motion of MurD, part of the peptidoglycan biosynthesis pathway in some bacteria, also demonstrates large scale

domain motion that was previously thought to be tied to ligand binding but may also be conformations the protein can access in the absence of ligand.[13]

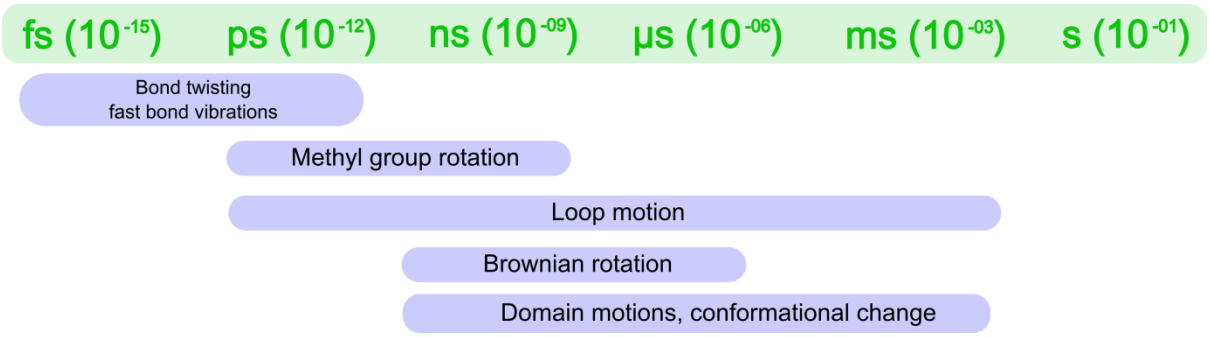


Figure 1.1: A timescale of the types of motion observed in proteins.

1.1.2 Methods used to study protein dynamics

The biophysical study of protein dynamics is intricate and difficult. One of the most successful techniques uses nuclear magnetic resonance, or NMR, in a variety of ways to uncover how proteins move in solution. Backbone dynamics can be investigated using amide ^{15}N and ^{1}H order parameters as this provides information about interactions such as hydrogen bond formation, while spin-lock and Carr-Purcell-Meiboom-Gill methods can provide information about slower motion such as the movement of loops.[13, 14] Relaxation dispersion techniques are commonly used to study dynamics in proteins typically in the micro-millisecond timescale.[15] One recent example used ^{15}N and ^{1}H relaxation dispersion measurements to investigate the motion of a "dynamic mutant" of *E. coli* dihydrofolate reductase with different ligands bound, demonstrating that the two types of measurements can provide different information about protein motion.[16]

Molecular dynamics simulations (MD), a computational tool that meshes chemistry and physics to simulate the movement of proteins at the atomic level, has been used extensively to probe how proteins move and to predict changes in conformation and dynamics upon ligand binding.[17] In one recent example, cryo-electron microscopy (cryo-EM) and large-scale molecular dynamics were combined to provide all-atom models for the HIV-1 capsid in several morphologies, demonstrating how MD can complement static methods, such as cryo-EM or X-ray crystallography, to provide a more complete picture of how a protein functions.[18] Covariance analyses, such as statistical coupling analysis, has also been combined with MD to investigate how coupled residues can preserve dynamics through evolution.[19]

Fluorescence resonance energy transfer (FRET) has been used to study large-scale domain rearrangements. FRET utilises donor and acceptor molecules that have been covalently linked to the protein of interest, and, if the molecules are close enough to one another, upon excitation of the donor molecule, the acceptor molecule fluoresces, which enables the measurement of conformational dynamics and distances in the protein.[20] Typically, this is achieved using an averaged population which presents difficulty in interpretation, but single molecule FRET, where the FRET signal of one molecule can be resolved, has been developed to overcome the problems of ensemble FRET.[20]

1.1.3 Dynamics and drug development

The study of protein dynamics is crucial for the development of new drugs. Mauldin et al.[21] used NMR spectroscopy to demonstrate that the binding of two inhibitors to dihydrofolate reductase altered the dynamics of a region distant to the active site, where the cofactor, NADH, binds. This prevented the interchange between NADH and NAD^+, and thus disrupted enzyme catalysis. This demonstrates how drug binding may have wide ranging impacts on a protein that are not easily detectable by techniques that directly investigate drug binding, such as fragment-based lead discovery or isothermal titration calorimetry (ITC), a point echoed by Peng et al.[22] who suggested that flexibility-function studies may provide new opportunities for drug design.

The role of protein dynamics in drug resistance is also becoming increasingly apparent. Podust et al.[23] determined the structure of *Mycobacterium tuberculosis* cytochrome P450 14α-sterol demethylase (*Mtu*CYP51) with azole inhibitors bound. CYP51 is an anti-fungal drug target, but mutations in this protein in clinically relevant species such as *Candida albicans* have led to drug resistance. In the above study, the authors mapped mutations from naturally occurring azole resistant strains from *C. albicans* onto the *Mtu*CYP51 structure and showed that the mutations that conferred resistance to the drugs were in regions that allowed CYP51 to move through its dynamic catalytic cycle. Drug resistance in HIV protease also developed at sites distant from the inhibitor binding site that affect protein dynamics.[24] Mutations conferring resistance occur in a region that is important for the large-scale conformational changes of the catalytic cycle of the HIV protease, and these mutations disrupt the motion of the enzyme, favouring the open and un-liganded conformation. This increases the off-rate of a competitive inhibitor, allowing the substrate sufficient time to bind and thus avoiding inhibition.

1.1.4 The evolution of protein dynamics

Protein dynamics are also critical for the evolution of new functions. One example of the essentiality of the evolution of dynamics is shown by the somatic evolution of antibodies. Affinity maturation, a process of somatic hypermutation and clonal selection, produces antibodies with increased affinity for an antigen.[25] Zimmerman et al.[26] explored the evolution from germ-line to mature antibodies in terms of the dynamics of the antibodies themselves. Mature antibodies tended to have mutations that induced hydrogen-bonding as well as packing interactions in particular locations. This leads to the rigidification of the combining site aiding the specificity and affinity of the antibody binding to the antigen. Adhikary et al.[27] also investigated the dynamics of antibodies, specifically in terms of their evolution. They showed that there were similarities in dynamics in most of the mature antibodies that came from different germ-line precursors, suggesting that altering the dynamics of the antibody is key for their evolution to be more specific, and to show more affinity, for a particular antigen. Additionally, a pair of antibodies from the same germ-line precursor produced one flexible and one significantly rigid pair of antibodies, and the rigid antibody bound the antigen with considerably higher affinity.

Another example of how protein dynamics can tune the evolution of new functions can be found in malate dehydrogenase and lactate dehydrogenase from Apicomplexa.[28] Although malate dehydrogenase and lactate dehydrogenase in Apicomplexa are similar in structure and catalyse the same type of chemistry, there is strict substrate specificity in their respective active sites. Through ancestral protein reconstruction and structural characterisation, it was shown that protein dynamics were key in evolution of malate dehydrogenase function from that of lactate dehydrogenase, as epistatic mutations far from the active site altered the dynamics of the active site significantly enough to alter substrate specificity.

1.2 Allosteric regulation

Allostery is simply defined as the process by which ligand binding at one site affects the function at a distant site.[29] Allosteric regulation is found in all types of self-replicating organisms.[30] Allosteric regulation of enzyme function by products of a metabolic pathway can commonly be found at control points. These points are typically near the start of a pathway, and can provide feedback regulation to the pathway.[31] Regulation of metabolism is key to organismal survival but flux-balance analysis, a computational approach, suggests that near-optimal survival bacterial growth can be achieved by simple product feedback inhibition or activation.[32] This demonstrates that allosteric control by feedback inhibition is both key to organismal survival, and part of a fascinating evolutionary process.

Three ways proteins can develop allosteric regulation have been proposed.[30] The first involves utilisation of loops or other dynamic parts of a protein that can form new binding sites for allosteric effectors. This was demonstrated by Mathonet et al.[33] where allosteric regulation by transition metal ions was introduced into an unregulated monomeric protein by insertion of random peptides into flexible loops followed by rounds of selection. The second is the formation of multi-subunit proteins, which allows communication between subunits and the potential for formation of binding sites for allosteric effectors.[30] The third way that proteins may evolve allosteric regulation is by domain swapping, caused by gene fusion events, where an unregulated enzyme recruits a domain that can bind a small molecule and interact with the catalytic domain.[34] Allosteric control can be introduced to unregulated enzymes by domain fusion events.[35, 36]

Guntas et al.[37] demonstrated the ease by which an allosteric enzyme could form by domain insertion of TEM-1 β-lactamase fragments into *E. coli* maltose binding protein (MBP), in which the β-lactamase activity was modulated by the ligand of the MBP protein. A domain insertion library was constructed in which a fragment of the lactamase gene was randomly inserted into a vector containing the *malE* gene, and selection was performed using an MBP auxotroph *E. coli* strain. Bi-functionality, followed by allostery, was then assessed. There were several allosteric enzymes identified, for example, an essentially end-to-end fusion that exhibited approximately 50% increased lactamase activity in the presence of maltose, as well as several mutants where the insertion had occurred in a so-called 'hot spot' for domain insertion.

1.2.1 The evolution of allosteric regulation

The evolution and conservation of allosteric regulation is a topic of hot debate. Allosteric mechanisms across a protein family may or may not be conserved depending on the protein family. Flock et al.[38] determined that, although there has been significant divergence in the GPCRs and their associated Gα proteins, there is a conserved mechanism of allostery that GPCRs utilise, and the authors identified key residues for allostery. A mutant analysis of Langerin, a C-type lectin receptor, discovered a network of evolutionarily conserved residues that are associated with Ca^{2+} binding, an essential co-factor that binds allosterically in a pH-dependent manner.[39] A computational approach looking at three bacterial CheY structures determined there was both conservation and variability in the allosteric response.[40] Indeed, although the *E. coli* and *Salmonella typhimurium* CheY sequences and structures share close similarity, there is a substantial difference in the identified hydrogen-bond network, suggesting that the underlying thermodynamic mechanism of allostery has evolved differently in the two examples. Clearly, further research is needed to understand the mechanisms of allostery in a variety of systems, especially with the advent of computational tools and the availability of sequence information.

1.2.2 Allostery and dynamics

Allosteric regulation and protein dynamics are intimately entwined. There are numerous examples of proteins that have altered dynamics in the presence of an allosteric ligand, sometimes in the absence of a conformational change. Tsai et al.[29] suggests that there are three types of allostery: one governed by entropy, one by both enthalpy and entropy, and one by predominantly by enthalpy, and the different types of allostery can produce different dynamics.

The first type of dynamics-driven allostery defined by Tsai et al.[29] is entropy-driven and is tied to a change in the overall dynamics of the system, such as increased rigidity or flexibility. There are numerous examples of purely entropy-driven allostery, but one key example is the binding of cAMP (cyclic AMP) to catabolite activator protein (CAP).[41] CAP is dimeric with two cAMP active sites, and the binding of cAMP to one active site alters the binding of a second cAMP to the other active site, an example of negative cooperativity. Two cAMP molecules bound to CAP increase the affinity of the protein to DNA, thus acting as allosteric effectors.[41] Popovych et al.[41] demonstrated that the negative cooperativity demonstrated by the binding of cAMP was solely as the result of changes in conformational entropy, namely the rigidification of CAP that negatively affects the binding of cAMP at the other binding site.

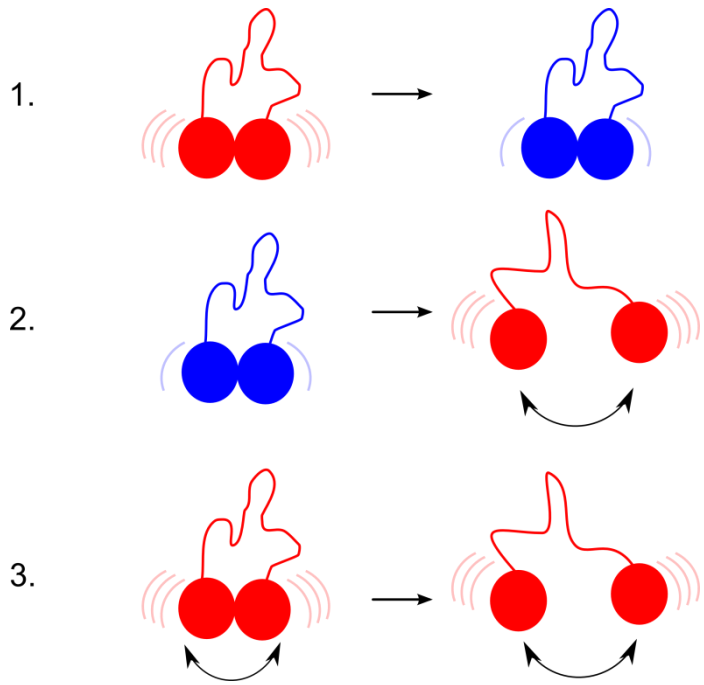


Figure 1.2: The three broad categories of changes in dynamics caused by allosteric ligand binding. 1. represents an overall change in dynamics without conformational change, while 2. represents changes in both dynamics and conformation, and 3. represents only a change in conformation

The second type of dynamic allostery combines both largely enthalpy-driven conformational change and changes to the intrinsic dynamics of the system that are entropically driven. One interesting example of this type of allostery is protein kinase A.[42] This protein exists in three major conformations depending on nucleotide binding. The overall energy landscape of protein kinase A is fairly broad, while the binding of allosteric inhibitors locks the protein into defined conformations, demonstrating changes in both wide-scale conformation and local dynamics.

The third type of allostery is characterised by large scale conformational change that alters the conformation of or access to the active site. One example is mammalian glutamate dehydrogenase (GDH) where the allosteric inhibitor GTP binds near a hinge region that separates the cofactor and substrate binding domains and traps the enzyme in a conformation that prevents opening of the catalytic cleft.[43]

Dynamically-driven allostery shows that allostery may still be occurring, even though a conformational change may not be detectable by methods such as X-ray crystallography.[29] Indeed, Gunasekaran et al..[44] argue that all dynamic proteins may be allosteric. This underlies the growing importance of dynamics in understanding allosteric regulation, and also how little is known about

how changes in conformational states can alter catalytic activity, especially upon binding of allosteric ligands at a distance from the active site.

1.2.3 Allosteric sites as drug targets

Conventionally, the active site has been the primary target of drug design. These compounds are termed orthosteric drugs. However, allosteric sites have also been proposed as drug targets. There are some advantages to allosteric drugs over orthosteric drugs. One key area in research into allosteric drugs is G-protein coupled receptors.[45, 46] As allosteric sites are typically less conserved than the active site, this can aid in the specificity of drug design to target a particular protein family.[47] Two types of G-protein coupled receptors, M_2 and M_3, showed a response to a particular drug, either inhibitory (M_3) or activating (M_2). A third type of receptor, M4, does not show an allosteric response even though it was shown that the receptor was binding the drug.[47] This demonstrates how one allosteric drug can produce a different response depending on the target. A study into the G-protein coupled receptor GLP-1 also showed what is termed ligand-induced stimulus bias, where different ligands binding to the receptor cause the receptor to adopt different conformations, and therefore produce different downstream signalling profiles.[48] The difference in response to the same drug shows how allosteric drugs could be used to mediate a specific downstream response. Allosteric sites can also be used to alter the activity at the active site, while orthosteric drugs typically abolish catalytic activity.[49] Allosteric and orthosteric drugs can also be used in combination to elicit the response required.[50]

There are several disadvantages to allosteric drug design, both biological and technical. As mentioned above, allosteric sites may be subjected to increased evolved resistance as the allosteric sites are typically less conserved. Also, small chemical differences in allosteric ligands can cause a major difference in allosteric response and subsequent downstream signalling, as found in the mGluR G-protein coupled receptors, where different allosteric drugs cause major differences downstream, complicating allosteric drug design.[51] Additionally, allosteric sites can be shallow, which provides challenges for the rational design of inhibitors due to problems with binding affinity of the allosteric ligand.[49]

Although there are significant issues to overcome in the development of allosteric inhibitors, they provide a new avenue to pursue in the development of new drugs. One particularly interesting example is the allosteric inhibition of the penicillin-binding protein 2 (PBP2a) from methicillin-resistant *Staphylococcus aureus* (MRSA).[52] This enzyme provides antibiotic resistance to β-lactam antibiotics (such as methicillin) to MRSA, and it was discovered that it is also allosteric, as

peptidoglycan binds at a site distant from the active site, that allows opening of the active site to facilitate substrate binding.[53, 54] *In silico* docking studies identified a class of compounds, the quinazolinones, that have potent activity against MRSA.[55] Further investigation into allosteric drug targets may provide additional weapons to fight the growing threat of antimicrobial resistance.

Although there is an intimate link between protein dynamics in various forms and allosteric regulation, there is a lack of knowledge about how such allosteric pathways have evolved, and how dynamics can drive evolution in different types of systems under varied evolutionary pressures. If allosteric drugs are going to provide an alternative to traditional orthosteric drugs, more research is needed to understand the evolution of protein dynamics, especially in relation to allosteric regulation. One such enzyme, where allostery appears to be driven by dynamics in the absence of a conformational change, is α-isopropylmalate synthase.

1.3 Isopropylmalate synthase (IPMS)

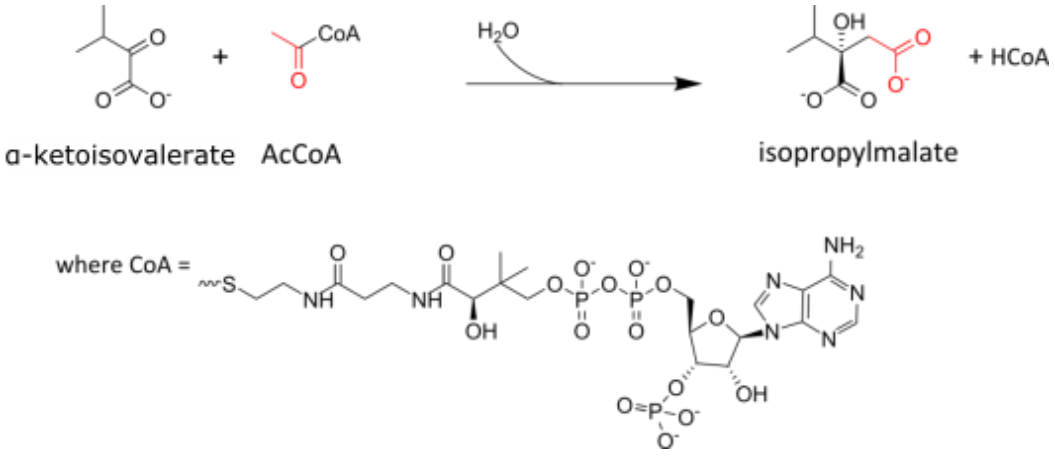


Figure 1.3: The chemical reaction catalysed by IPMS.

α-Isopropylmalate synthase (IPMS) catalyses the first committed step of leucine biosynthesis in microorganisms, utilising ketoisovalerate (KIV) and acetyl co-enzyme A (AcCoA) to form isopropylmalate (IPM) and CoA (Figure 1.3). The leucine biosynthetic pathway is essential to the proliferation of microorganisms, and its absence in higher eukaryotes, makes IPMS is a promising target in the search for new antibiotics.[56] Leucine auxotrophy in *Mycobacterium tuberculosis* (*Mtu*) reduces virulence and, when immune deficient mice have been infected with a leucine auxotroph strain of *M. tuberculosis*, this infection provides protection against subsequent infection.[57] As with most pathogens, macrophages engulf *Mycobacteria* upon infection. *Mycobacteria* can then alter the environment of the phagosome, allowing for bacterial replication inside the macrophage.[58] Bange et al.[59] determined that a leucine auxotrophic *Mycobacterium bovis* BCG strain could not replicate inside cultured macrophages, showing why this strain could not grow within mice. The essentiality of leucine biosynthesis has also been demonstrated in *E. coli*, and in *Methanococcus maripaludis*, where it was shown that leucine auxotrophy can be produced by knocking out the *leuA* gene that codes for IPMS.[60, 61] IPMS has thus been proposed as a putative drug target, especially in *M. tuberculosis*.[62]

IPMS catalyses a Claisen-like condensation reaction between KIV and AcCoA to produce isopropylmalate (Figure 1.3).[63] All IPMS enzymes characterised thus far require a divalent metal ion for catalysis.[64] As with many metabolic pathways, as IPMS catalyses the first committed step of the pathway, it is feedback inhibited by L-leucine, the end-product of the pathway. IPMS is allosterically regulated by L-leucine binding to the C-terminal regulatory domain, inhibiting catalysis. IPMS is also inhibited by CoA, a product of the reaction, in organisms such as

Saccharomyces cerevisiae.[65] The gene that encodes IPMS, *leuA*, is also under transcriptional control by L-leucine, as demonstrated by experiments in both *E. coli* and *Salmonella typhimurium.*[66]

1.3.1 The crystal structure of *Mycobacterium tuberculosis* IPMS

The crystal structure of IPMS from *M. tuberculosis* (*Mtu*IPMS) has been solved, and this provides a basis for further study of this type of enzyme.[67] However, this is the only full-length IPMS structure that has been solved, and there are significant phylogenetic differences between the 'IPMS2' (*Mtu*IPMS-like IPMSs) and the 'IPMS1' (*Nme*IPMS-like IPMSs) groups.[68, 69] There may be differences in dynamics between the two groups that are not apparent from the overall structure.

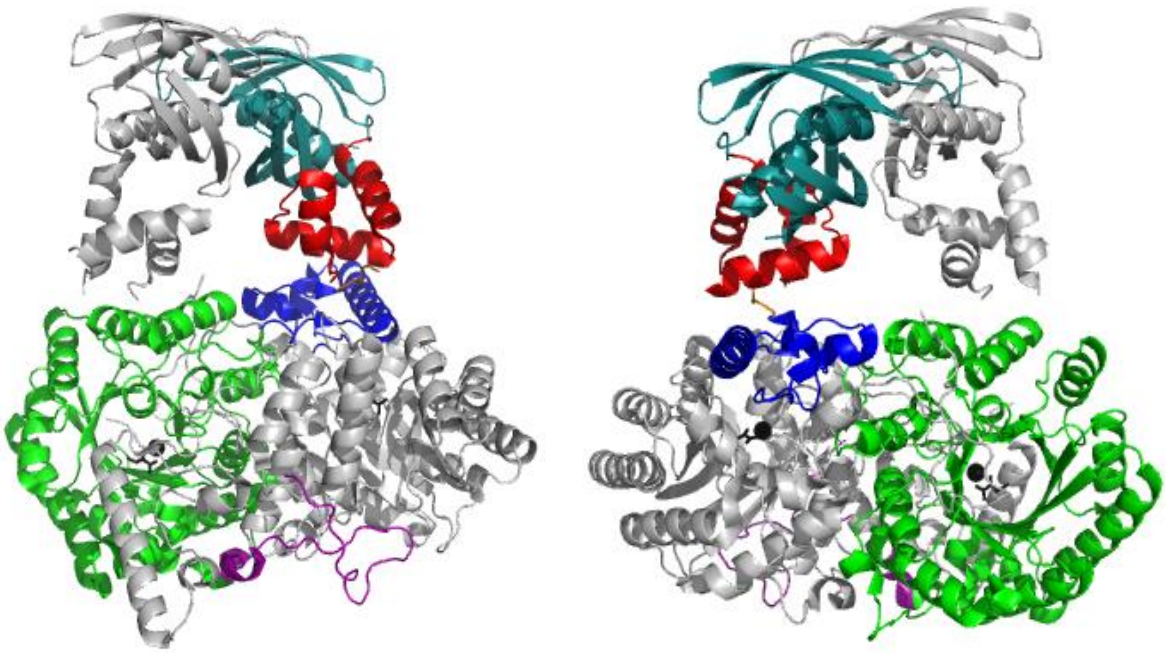


Figure 1.4: The structure of *Mtu*IPMS (PDB: 1SR9). The N-terminal extension is shown in purple, the catalytic domain in green, subdomain I in blue, subdomain II in red, and the regulatory domain in teal in Chain B. The essential metal ion (spheres) and the substrate, KIV (stick), are shown in are shown in black. Chain A is shown in grey. The protein is domain-swapped and is rotated 180°.

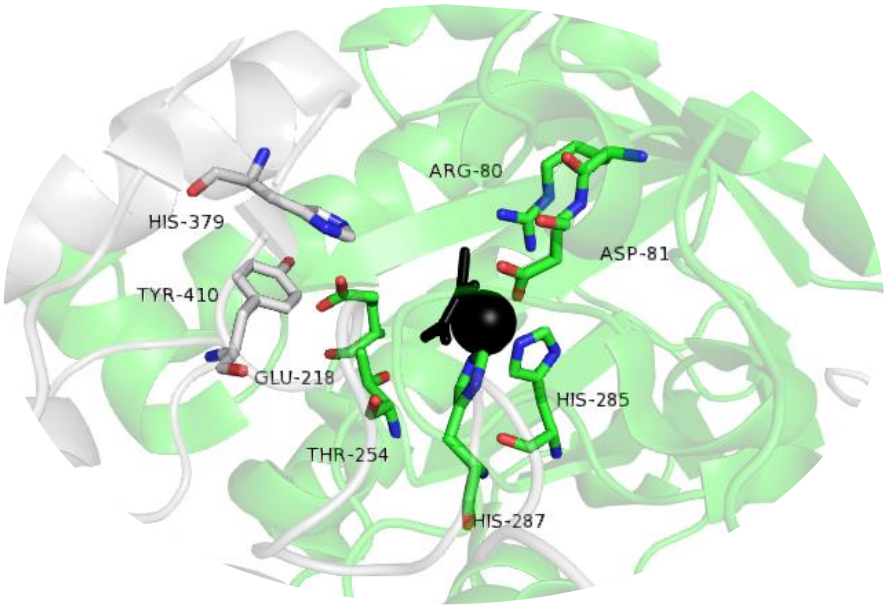


Figure 1.5: KIV binding in *Mtu*IPMS (PDB: 1SR9). The metal ion and KIV are shown in black.

*Mtu*IPMS is a domain-swapped homodimer with the subunit structure comprised of three main components: The N-terminal catalytic domain, which is a triosephosphate isomerase (TIM) $(\beta/\alpha)_8$-barrel, two subdomains (subdomains I and II) that form a catalytic accessory unit between the catalytic domain, and the C-terminal regulatory domain (Figure 1.4).[67] Structures of *Mtu*IPMS with bound substrate KIV (PDB: 1SR9) and inhibitor L-leucine (PDB: 3FIG) have been solved, as well as several with other non-natural ligands bound in the active site. As of yet, no structure of an IPMS has been solved with either CoA or AcCoA bound, nor has the apo structure of *Mtu*IPMS been solved. The catalytic domain of *Mtu*IPMS also has an N-terminal extension that contributes to dimerization.[67] The active site, as with many TIM barrels, is at the C-terminal end of the barrel.[70] C-terminal to the catalytic domain is subdomain I, the first part of the catalytic accessory unit. Subdomain I is formed from one long and two short α-helices. The long helix sits across over the active site of the other chain and contributes residues to the active site of the opposite chain. Subdomain II, the second part of the catalytic accessory unit, is composed of three α-helices that form a tight bundle and is connected to subdomain I by a flexible linker that is disordered in both the L-leucine bound and KIV bound structures of *Mtu*IPMS. Subdomain II also forms a close interaction with the C-terminal regulatory domain, that is formed of a novel $(\beta\beta\beta\alpha)_2$ fold, where L-leucine binds in the interface of the regulatory domain dimer. In the active site, the essential metal ion is co-ordinated by two conserved histidine residues (His285 and His287) and an aspartic acid (Asp81) from the conserved LR(D/E)G motif (Figure 1.5). KIV forms interactions with Arg80 and Thr254 as well as with the metal ion. Glu218, and Tyr410 and His379 from the other

chain also contribute to the active site, and are strongly or completely conserved in IPMS.[67] In related enzymes, the smaller substrate binds first followed by AcCoA in an ordered fashion, and this is likely to occur in IPMS as well.[71] There is some capacity in the *Mtu*IPMS active site to bind alternative substrates – α-ketobutyrate, α-ketovalerate, and pyruvate have been shown to act as poor substrates for the condensation reaction[72]. Unlike the smaller substrate, *Mtu*IPMS shows specificity for AcCoA as the larger substrate.[56]

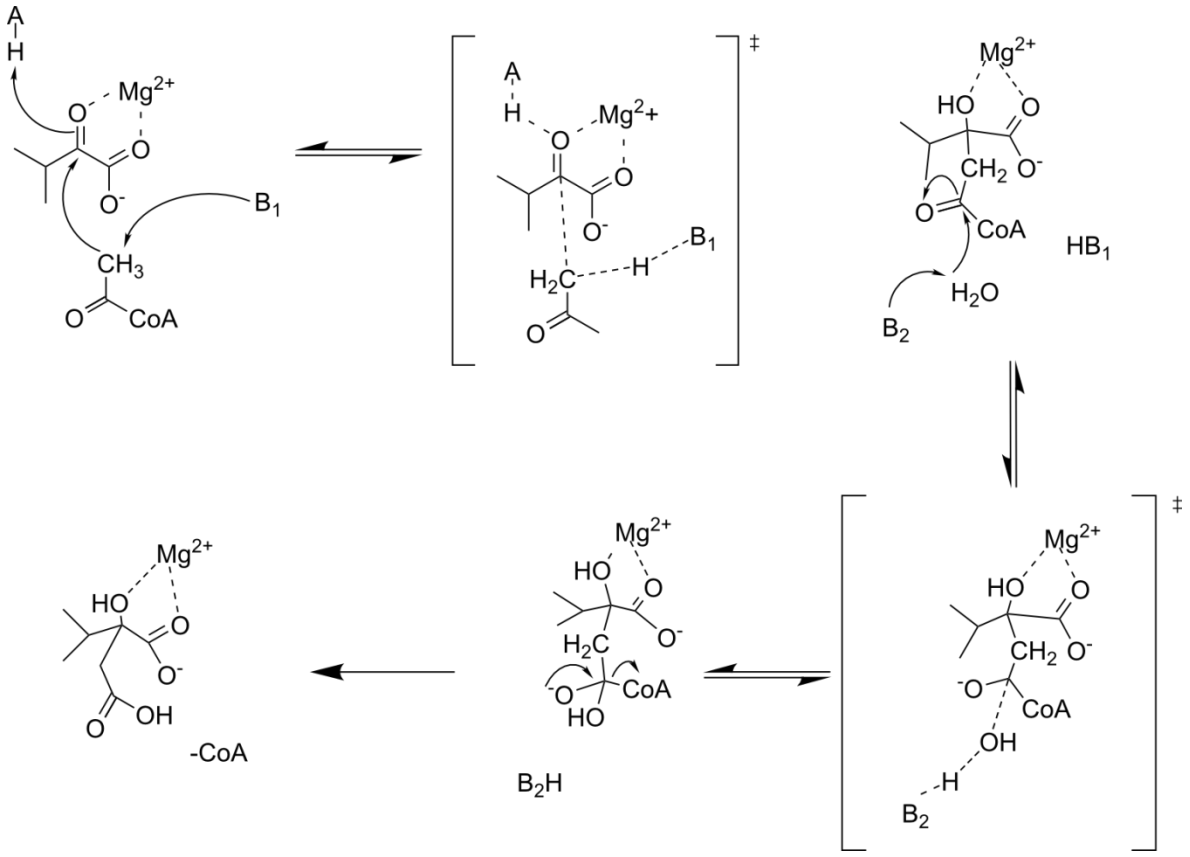


Figure 1.6: The kinetic mechanism of *Mtu*IPMS as determined by de Carvalho et al..[56] A, B_1 and B_2 are acidic (A) and basic (B) residues in the enzyme active site that are involved in the reaction.

The kinetic mechanism of *Mtu*IPMS has been studied in detail using a variety of techniques (Figure 1.6).[56] These experiments suggested that *Mtu*IPMS uses a non-rapid equilibrium, random, bi-bi kinetic mechanism and that there are likely two catalytic bases - one that deprotonates and enolises AcCoA and another that hydrolyses the isopropyl-CoA intermediate into the two products. Interestingly, solvent isotope labelling experiments suggest that the chemistry of the reaction is not the rate-limiting step, and further study suggests that product release may be the rate determining step in *Mtu*IPMS.[56, 73]

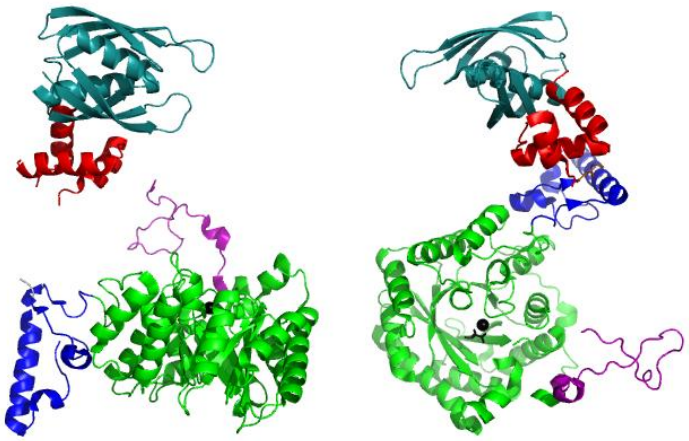


Figure 1.7: The structure of *Mtu*IPMS (PDB: 1SR9).Chain A (left) and Chain B (right) showing the structural asymmetry of the dimer. The N-terminal extension is shown in purple, the catalytic domain is shown in green, subdomain I is shown in blue, subdomain II is shown in red, and the regulatory domain is shown in teal

The crystal structure of *Mtu*IPMS is asymmetric, meaning that the two chains superimpose poorly (Figure 1.7).[67] The linker region between subdomains I and II was not resolved in any crystal structure of *Mtu*IPMS thus solved. This asymmetry may play important roles in catalysis and the allosteric communication pathway between the regulatory domain and the active site.[67] It is not clear whether the asymmetry is due in part to crystal packing constraints, although it is a conformation that is accessible by the protein.[74]

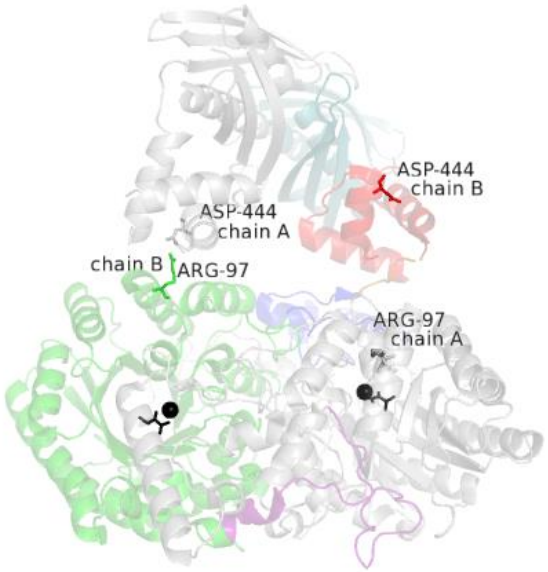


Figure 1.8: The crystal structure of *Mtu*IPMS (PDB: 1SR9) highlighting residues Arg97 and Asp444.

Point mutations in *Mtu*IPMS at Arg97 and Asp444, located in the catalytic domain and subdomain II respectively, were made to investigate the domain-swapped dimer interfaces (Figure 1.8).[74] These

protein variants both adopted similar solution structures to that of wild-type *Mtu*IPMS in the presence and absence of L-leucine, as determined by small angle x-ray scattering (SAXS), but had increased catalytic activity and decreased sensitivity to L-leucine compared to the wild-type protein. It was suggested that these mutations increased the flexibility at one interface formed by the catalytic domain of one chain and subdomain II from the other chain. In the other half of the structure, the two residues are approximately 31 Å apart. The increase in flexibility at this interface may allow regions important for AcCoA interaction and binding to sample conformations that allow AcCoA to bind more readily, thus showing increased catalytic activity compared to the wild type protein. The decrease in inhibition by L-leucine may also occur due to the increase in flexibility at the interface in this mutant protein.[74]

1.3.2 Structures of other IPMSs

Several partial structures of IPMS enzymes have been solved, including the IPMS from *Neisseria meningitidis* (*Nme*IPMS) (Figure 1.9) truncated at residue Glu365 in subdomain II (PDB 3RMJ), and *Mtu*IPMS, which was truncated at residue Val425 in subdomain II (PDB 3U6W).[64] The truncated enzymes had structurally similar catalytic barrels, both to each other and to the respective full length protein, but the structures of the partial subdomains showed substantially more flexibility than that of the full-length protein. Neither of these truncated proteins was catalytically active, although it was shown that both truncated enzymes could bind KIV.[64]

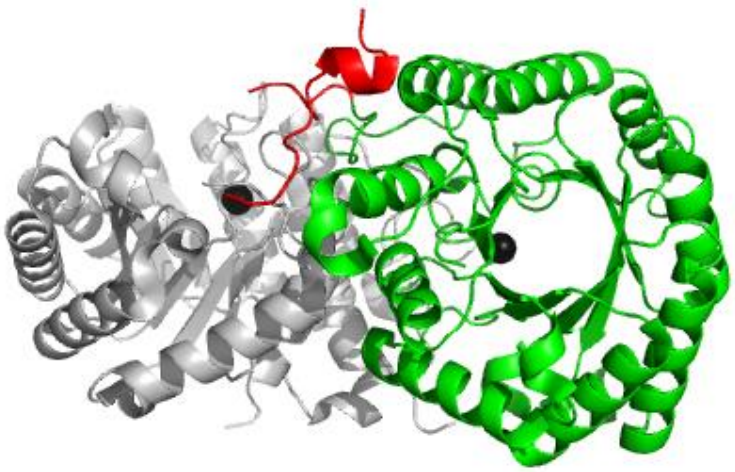


Figure 1.9: The structure of a truncation of *Nme*IPMS (PDB: 3RMJ).The protein was truncated at position Glu365. The metal ion is shown in black, denoting the location of the active site, while the catalytic domain is shown in green and part of subdomain I is shown in red

A full length, but naturally truncated, IPMS structure has also been solved (Figure 1.10).[2] The 'short form' IPMS from *Leptospira biflexa* (*Lbi*IPMS2) does not possess the C-terminal regulatory domain, but maintains catalytic activity, although not allosteric regulation by L-leucine. Structurally, the TIM barrel is similar to that of *Mtu*IPMS, with KIV bound in a similar position, making the same contacts seen in the *Mtu*IPMS active site. Subdomain II adopts a different conformation in *Lbi*IPMS2 compared to subdomain II in *Mtu*IPMS. Unlike the asymmetry seen in *Mtu*IPMS, the two chains in *Lbi*IPMS2 are symmetrical in the crystal structure. Subdomain II appears to adopt a similar conformation to that of Chain A of the *Mtu*IPMS structure but is substantially different to that of Chain B of *Mtu*IPMS. Truncations of subdomain II in both the long form (*Lbi*IPMS1) and short form (*Lbi*IPMS2) enzymes demonstrated that while the absent C-terminal regulatory domain is not required for catalysis, an intact subdomain II is.[2]

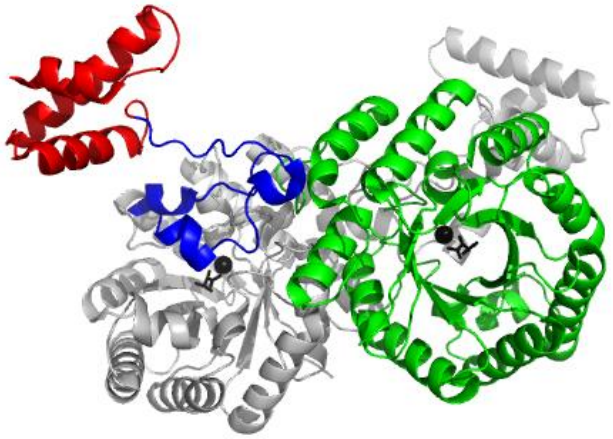


Figure 1.10: The structure of *Lbi*IPMS2 (PDB: 4OV4).One chain is shown in grey. The catalytic domain is shown in green, subdomain I is shown in blue, and subdomain II is shown in red.

*Lbi*IPMS1 and *Lbi*IPMS2 also showed cooperativity with regards to the substrates. *Lbi*IPMS2 demonstrated positive cooperativity towards both KIV and AcCoA, although the cooperativity was much more pronounced with AcCoA than KIV.[2] Positive cooperativity was less pronounced in the long form *Lbi*IPMS1 for which no structure has yet been solved. Cooperativity has also been seen in substrate binding in other IPMSs although it has not been detected in *Mtu*IPMS.[75, 76]

1.3.3 Allosteric regulation of IPMSs

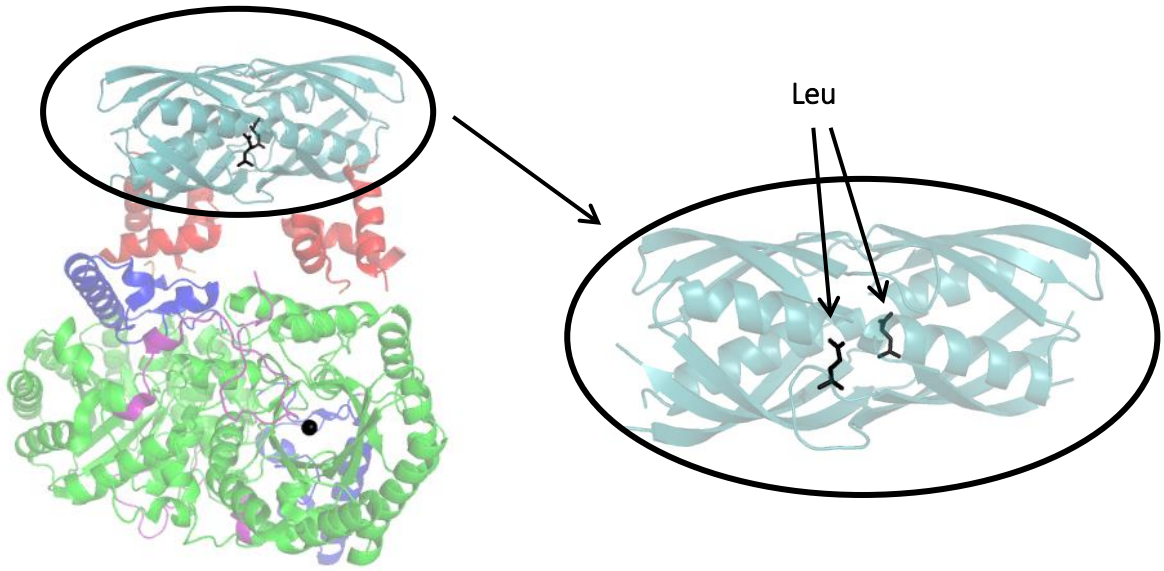


Figure 1.11: The position of L-leucine bound to the regulatory domain of *Mtu*IPMS (PDB 3FIG).The N-terminal extension is shown in purple, catalytic domain is shown in green, subdomain I is shown in blue, subdomain II is shown in red, and the regulatory domain is shown in teal. The black sphere denotes zinc bound in the active site while L-leucine is shown bound in the regulatory domain as black sticks.

As mentioned above, IPMS is allosterically regulated by the end product of the pathway, L-leucine. L-Leucine binds at the dimer interface between the regulatory domains (Figure 1.11) and thus mediates catalysis at the active site, with two molecules of L-leucine bound per homodimer. Interestingly, crystal structures of both apo and L-leucine-bound *Mtu*IPMS have the same overall conformation, including the structural asymmetry discussed previously, which suggests that L-leucine binding may alter the molecular dynamics of the IPMS protein, rather than cause a discrete conformational change upon binding.[67] However, it is also plausible that there are more localised conformational changes that are not revealed by X-ray crystallography due to effects such as crystal packing.[77] Small-angle X-ray scattering data of *Mtu*IPMS also suggest that there is no gross conformational change of the protein in solution upon L-leucine binding.[74]

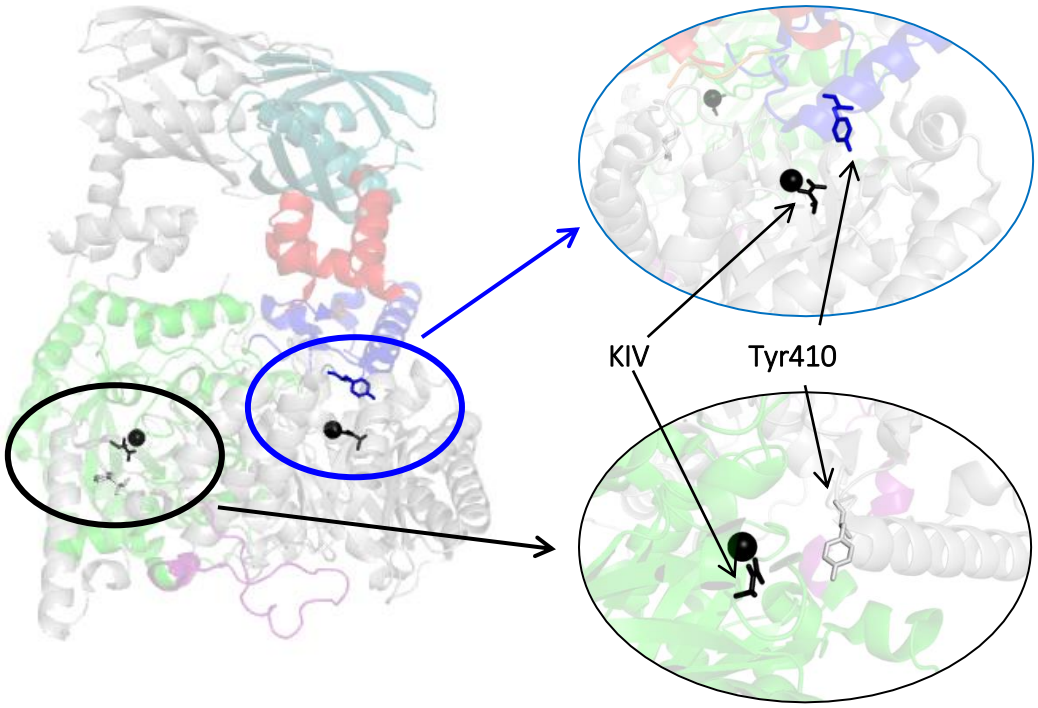


Figure 1.12: The structure of *Mtu*IPMS (PDB: 1SR9) highlighting the location of Tyr410.Tyr410 is shown in Chain A (black circle) and chain B (blue circle). Chain A is shown in grey while the catalytic domain of Chain B is shown in green, subdomain I is shown in blue, subdomain II in red, and the regulatory domain is shown in teal. The active site is denoted by the substrate, KIV (black stick) and the metal ion (black sphere).

Although the crystal structures with and without L-leucine have been solved, there is very little difference between the two structures.[63] De Carvalho et al.[63] constructed point mutations in the region of the protein where subdomain I interacts with the catalytic barrel to investigate how allosteric inhibition by L-leucine functioned in the absence of conformational change. Residue Tyr410 (Figure 1.12), that resides in subdomain I and thus crosses over to form part of the active site with the other chain, was mutated to Phe, and the Tyr410Phe mutant was entirely insensitive to L-leucine even though this residue is far from the allosteric site. This result suggests that there is a communication network extending through the subdomains that is important for allosteric communication.

Although the precise mechanism of allostery remains elusive, it has been shown that dynamics are critically important to this enzyme for catalysis. Frantom et al.[78] utilised hydrogen/deuterium exchange (HDX) to explore the mechanism by which the allosteric signal is transferred from the regulatory domain to the catalytic domain. In wild-type *Mtu*IPMS, the comparison of HDX in the apo and L-leucine-bound *Mtu*IPMS showed a reduction in exchange in residues that surround the allosteric site. This technique identified a predominantly hydrophobic region in subdomain II that

interacts with the regulatory domain as part of an allosteric network that transmits the signal to the catalytic domain. There was also reduced exchange in part of the catalytic barrel that is important for KIV binding, suggesting a potential mechanism by which L-leucine binding affects catalysis.

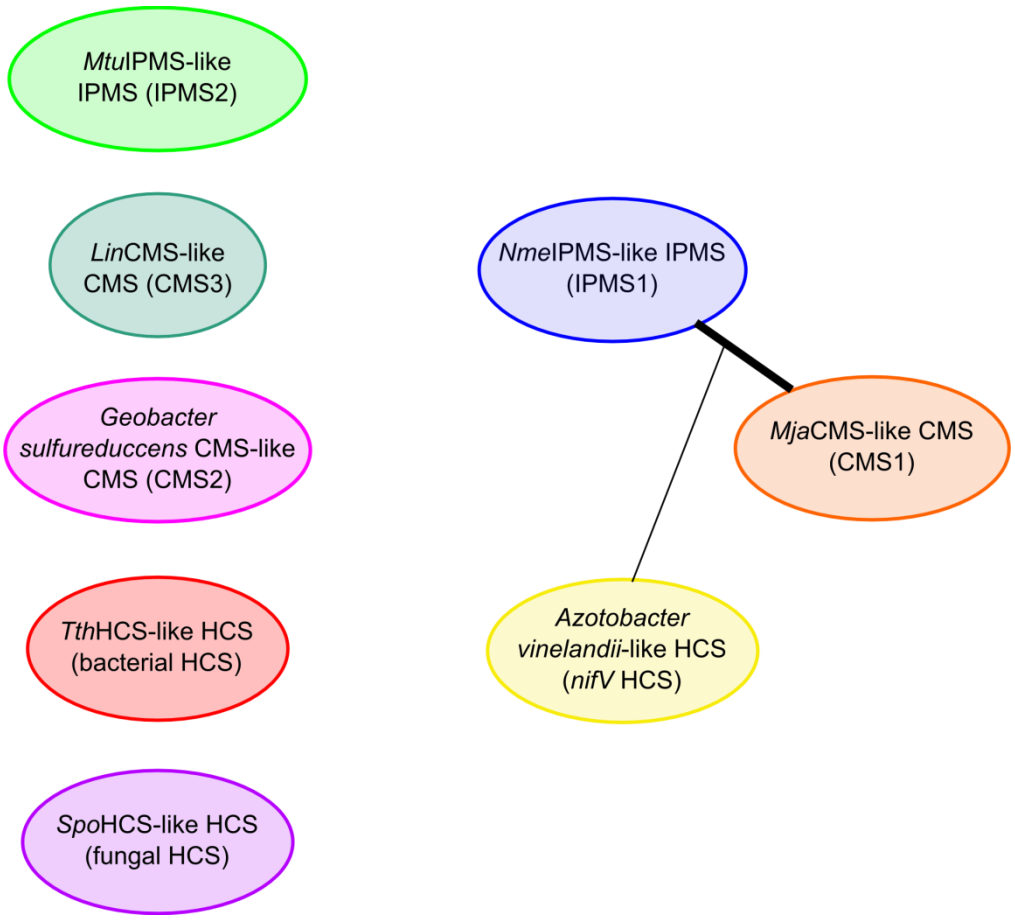


Figure 1.13: The diversity of the IPMS and IPMS-like protein family based on Kumar et al..[69] The colours indicate the different groups of proteins.

There appear to be different mechanisms of allostery in different IPMS enzymes. V-type allostery, where binding of the allosteric inhibitor affects the maximum activity, is seen in *Lbi*IPMS1 towards KIV upon binding of L-leucine, whereas a mixed V/K-type allostery is shown towards AcCoA.[2] K-type allostery occurs when binding of the inhibitor affects the affinity of the enzyme towards a substrate. *Mtu*IPMS has a different form of allosteric regulation, showing slow-onset inhibition that is not seen in other examples of IPMS.[79] Also, *Mtu*IPMS solely demonstrates V-type allostery towards both substrates.[73] *Nme*IPMS demonstrates mixed, non-competitive inhibition for both KIV and AcCoA.[1] The differences in mechanism of allosteric regulation towards the substrates may represent phylogenetic differences, as *Mtu*IPMS is phylogenetically distinct from the other

two IPMS enzymes (Figure 1.13).[69] However, Kumar et al.[80] showed that both *Methanococcus jannaschii IPMS* (*Mja*IPMS) , a IPMS1 IPMS, and *Mtu*IPMS, a IPMS2 IPMS, demonstrate V-type allostery that targets the hydrolytic step of the chemical mechanism. Kumar et al.[80] suggest that although they are phylogenetically different and only share ~20% sequence identity, the allosteric mechanism has been conserved or has shown convergent evolution. Dong et al.[8] argues that allosteric mechanisms can be transmitted through multiple, different pathways that pre-exist in the protein. The binding of an allosteric inhibitor creates strain in the dynamic protein, and this strain dissipates through these extended pathways, altering the ensemble population, producing different effects. The difference in mechanism by which allostery occurs, and that this is not tied directly to phylogeny, suggests that the different enzymes may individually utilise different pre-existing pathways to facilitate allosteric regulation.

The dynamics of IPMS are intricate and show significant differences between individual proteins. The dynamics of IPMSs are tied to both the enzyme activity and allostery. This group of enzymes present a fascinating picture of how dynamics can be acted upon by evolutionary processes in both enzyme catalysis and in allosteric regulation.

1.4 Evolution of amino acid biosynthesis pathways

This project focuses on three amino acid biosynthetic pathways in prokaryotes: the leucine biosynthetic pathway, the threonine-independent isoleucine biosynthetic pathway (termed the isoleucine biosynthesis pathway in this thesis), and the α-aminoadipate pathway for lysine biosynthesis (the lysine biosynthesis pathway) (Figure 1.15). The leucine biosynthetic pathway is found in most Bacteria, Archaea, and Fungi, as well as green plants.[81] The threonine-independent pathway for isoleucine biosynthesis is only found in a subset of bacteria such as *Geobacter sulfurreducens* and *Leptospira interrogans* (*Lin*), and is common in Archaea.[82, 83] The α-aminoadipate pathway for lysine biosynthesis is common in Fungi, but a variation of this pathway is also found in *Thermus-Deinococcus* bacteria as well as some Archaea.[84, 85]

The first enzymes in the pathways detailed above, IPMS, homocitrate synthase (HCS), which is present in the lysine biosynthetic pathway, and citramalate synthase (CMS) in the isoleucine biosynthetic pathway, are structurally homologous to each other. Gene duplication, or horizontal gene transfer, and functional divergence of an ancestral, promiscuous, IPMS, may have given rise to these enzymes.[86-88] The leucine biosynthetic pathway is the only known pathway for the biosynthesis of L-leucine, suggesting that the ancestral protein was an IPMS, as there are alternative pathways for the synthesis of isoleucine and lysine that do not rely on enzymes with structural homology to an IPMS.[88]

Drevland et al.[89] suggests that the evolution of the pyruvate pathway for isoleucine biosynthesis occurred via gene duplication and divergence of the leucine biosynthesis pathway, and that this occurred at least twice, producing the *Leptospira interrogans* (*Lin*)-like CMSs (CMS3) and *Mja*-like CMSs (CMS1) identified by Kumar et al..[69] There also appear to be multiple origins of the IPMS from a zygomycete fungus, *Phycomyces blakesleeanus*, where the *leuA* gene is more closely related to plants and cyanobacterial *leuA* genes than it is to other fungal homologues.[90] In a study examining the effects of a feedback-insensitive IPMS in wild and cultivated tomatoes, it was also suggested that there were independent and lineage-specific gene duplication and diversification events from IPMS to produce either the methylthioalkylmalate synthase (MAM synthase), another paralogue of IPMS, in Brassicaceae and the feedback insensitive IPMS in tomato and related plants.[91] The similarity in substrate selectivity in these enzymes is demonstrated in Figure 1.14. This demonstrates another instance of gene duplication and divergence in the IPMS and IPMS-like group of enzymes.

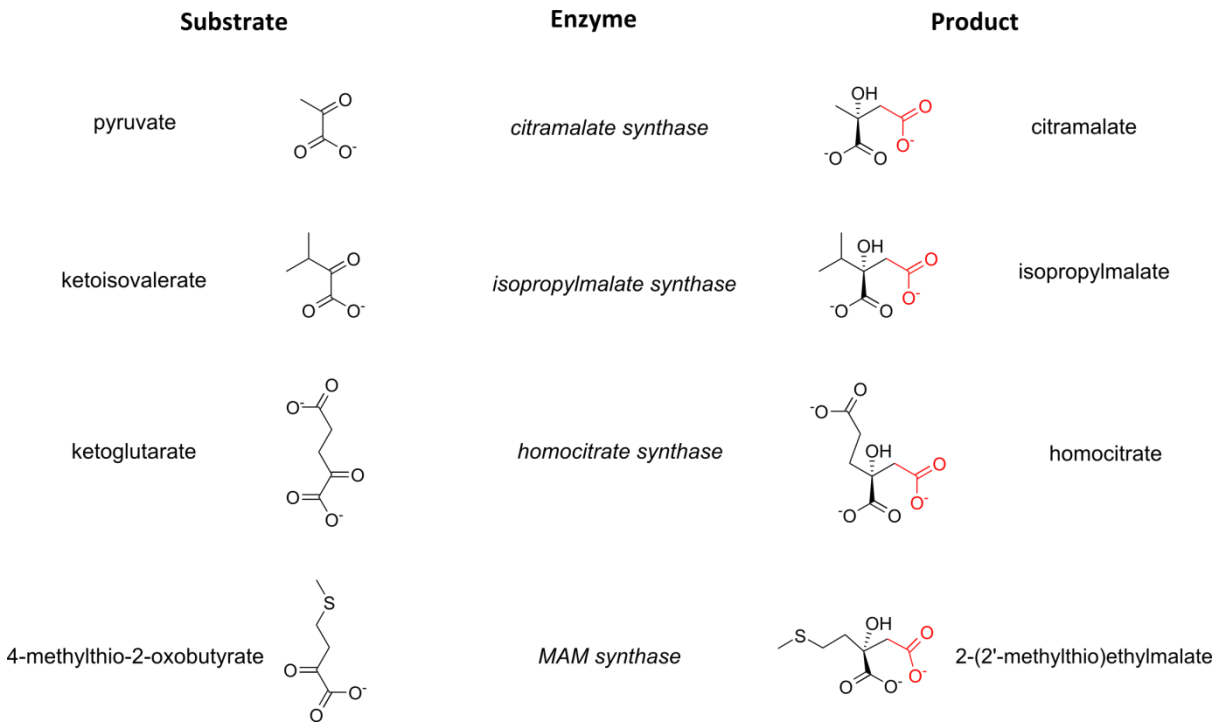


Figure 1.14: The similarity in substrate between IPMS and three homologues.All four enzymes are metalloenzymes and utilise AcCoA as well as the substrates detailed in this figure.

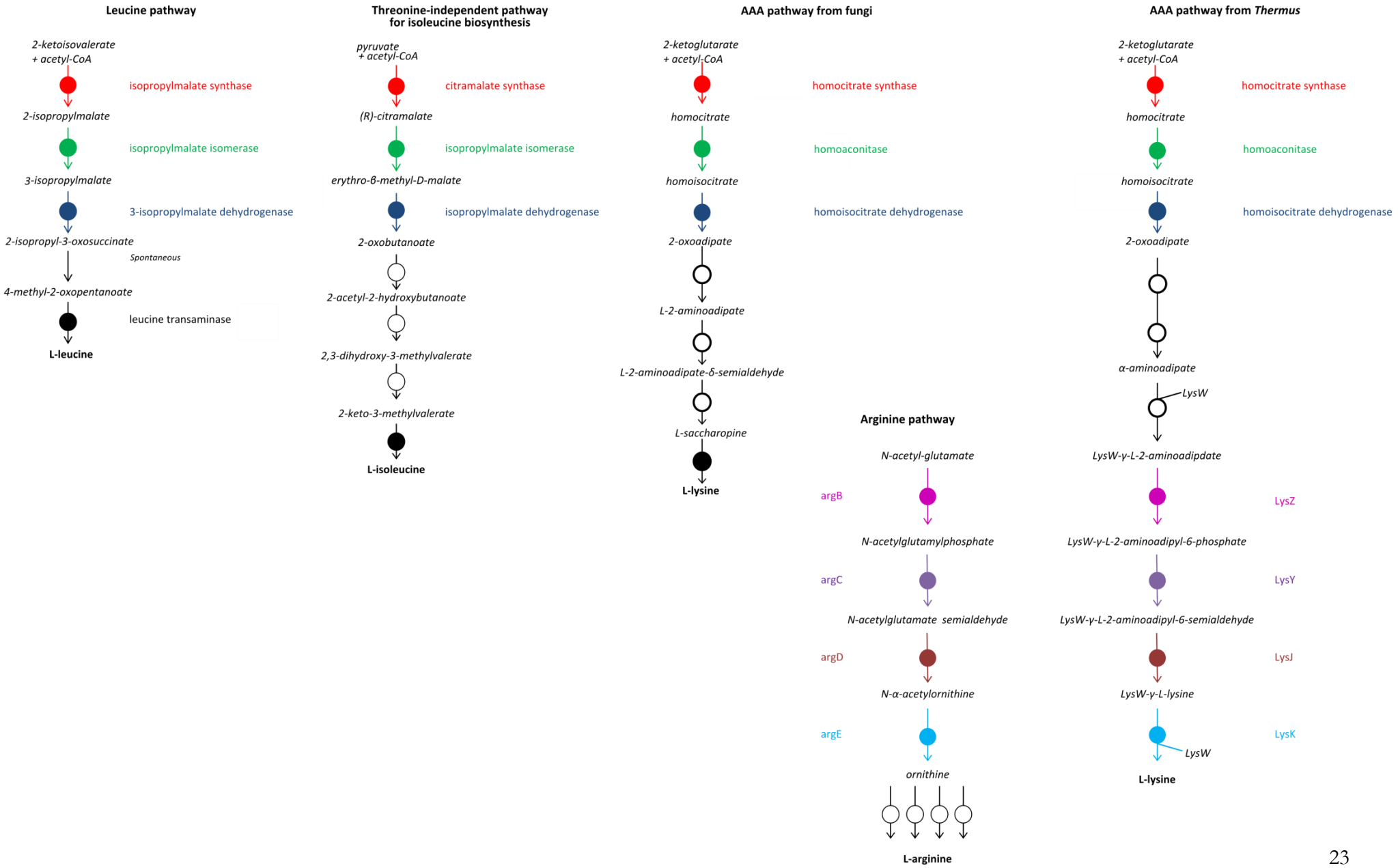


Figure 1.15: Biosynthetic pathways of interest from different organisms.The leucine biosynthetic pathway, the threonine-independent isoleucine biosynthetic pathway, and the two variants of the aminoadipate pathway for lysine biosynthesis from Fungi, and from thermophilic bacteria, showing the homology with the arginine biosynthesis pathway from bacteria. Colours represent homology (e.g. IPMS and IPMS-like proteins are in red. Whited/black circles represent where no homology is observed.

1.4.1 Substrate promiscuity in Claisen-condensation-like enzymes

It has been shown that there is some degree of substrate promiscuity in the IPMS and IPMS-like enzymes.[72] *Mtu*IPMS, as discussed above, can catalyse the condensation reaction using alternative ketoacid substrates, albeit poorly. Although *Mja*CMS appears to show no activity with alternative substrates, *Lin*CMS can utilise ketobutyrate and glyoxylate.[92, 93] Homocitrate synthase from *Thermus thermophilus* can use oxaloacetate as a substrate, instead of the natural substrate α-ketoglutarate, in the presence of potassium chloride.[94] This suggests that, although the active sites of the enzymes are tailored towards a specific substrate, there is some flexibility as to which substrate can be bound.

There is preliminary evidence for a bifunctional, or promiscuous, IPMS/HCS in *Pyrococcus horikoshii*, a thermophilic archaeon.[95] Only one copy of the genes coding for the first three steps in either lysine or leucine biosynthesis has been identified in the genome even though both lysine biosynthesis via the AAA pathway and leucine biosynthesis are thought to occur.[85] Interestingly, the other pathway for lysine biosynthesis, the L,L-diaminopimelate (DAP) pathway, is paralogous to an arginine biosynthesis pathway, thus suggesting a common ancestor for these pathways as well, demonstrating how a biosynthetic pathway can be built modularly from existing pathways.

In *Thermus* species, the first half of the pathway to produce aminoadipate is analogous to the fungal AAA pathway (Figure 1.15). From there, however, the pathways diverge. The fungal AAA pathway utilises two reductase enzymes and a dehydrogenase enzyme to form lysine, whereas in *Thermus* species, there has been a gene duplication event involving elements of the arginine biosynthesis pathway, and the latter half of the *Thermus* AAA lysine biosynthesis pathway shares common ancestry with arginine biosynthesis (Figure 1.15).[95] This was demonstrated by Miyazaki et al.[96] who characterised LysJ, the protein product of the *lysJ* gene, a homologue of *argD*, from *Thermus thermophilus*, that is essential for lysine biosynthesis. Fondi et al.[95] also demonstrated that the *Thermus* AAA pathway for lysine biosynthesis was only present in a small subset of organisms, namely *Sulfolobus* and *Pyrococcus* archaea and very few bacteria. The DAP pathway is well represented by other bacterial species, whereas the fungal AAA pathway is, as of yet, the only lysine pathway seen in fungi. This suggests there may be two potential evolutionary origins of the AAA pathway, utilising different gene duplication events.

Interestingly, homocitrate synthase also appears to have other roles outside of amino acid biosynthesis. It has been shown that homocitrate synthase, of which there are two isoforms coded by the *lys20* and *lys21* genes, in *Saccharomyces cerevisiae*, can be localised to the nucleus and has a role

in DNA damage repair.[97] This appears to be independent of the homocitrate synthase catalytic function that is present in both Lys20 and Lys21. It also appears that the Lys20 isoform of HCS in *S. cerevisiae* may have weak histone acetyltransferase (HAT) activity while also interacting with other HATs, namely Gnc5 and Esa1. Furthermore, homocitrate synthase and two other enzymes that catalyse the first three steps of the AAA lysine biosynthesis pathway have also been implicated in methane production in methanogenic archaea.[98] In the typical AAA pathway reaction, homocitrate synthase utilises α-ketoglutarate and acetyl CoA to produce (*R*)-homocitrate (Figure 1.15).[99] Homoaconitase then dehydrates homocitrate to produce *cis*-homoaconitate before homoaconitate hydratase activity of homoaconitase utilises the product to form homoisocitrate (Figure 1.15). Homoisocitrate dehydrogenase acts on homoisocitrate to form 2-oxoadipate. However, in the production of coenzyme B which is essential for methane production in methanogenic archaea, instead of conversion of oxoadipate to L-2-aminoadipdate as in AAA-pathway catalysed lysine biosynthesis, 2-oxoadipate is extended to form, finally, 2-oxosubterate, required to form the final product, 7-mercaptoheptanoylthreonine phosphate or coenzyme B.[100]

Homocitrate synthase has also been implicated in a third pathway, the production of an iron-molybdenum cofactor that plays a critical role in nitrogen fixation by bacteria. The cofactor contains one molecule of (*R*)-homocitrate produced by homocitrate synthase coded for by the so-called *nifV* gene in organisms such as *Azotobacter vinelandii*.[101] The diverse roles of homocitrate synthase in pathways beyond lysine biosynthesis in some organisms demonstrate how some enzymes can be utilised in vastly different pathways depending on the needs of the organism.

IPMS, CMS, and HCS are related through structure and catalytic activity, even though the primary amino acid sequences have diverged significantly. All three enzymes are key catalytic steps in three distinct pathways that have likely diverged at some point from an original common, promiscuous, ancestral pathway. Studying these three enzymes will provide more insight into the evolution of these pathways.

1.5 Citramalate synthase

Citramalate synthase (CMS) catalyses the first committed step in the synthesis of L-isoleucine in one of two pathways utilised by, primarily methanogenic, microorganisms.[89] CMS, like IPMS, utilises AcCoA as a substrate, but uses pyruvate as the α-ketoacid substrate to produce citramalate (Figure 1.16).

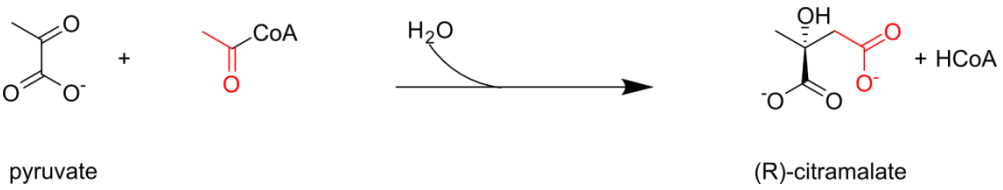


Figure 1.16: The reaction catalysed by citramalate synthase.

A full crystal structure of CMS has not yet been solved, but two partial structures of the catalytic domain and regulatory domain of CMS from *Leptospira interrogans* (*Lin*CMS), have been solved (Figure 1.17).[78, 92] *Lin*CMS is very similar in structure to *Mtu*IPMS, despite there being only approximately 30% sequence similarity between the two proteins. IPMS and CMS are the only characterised enzymes to bear the unique fold of the regulatory domain.

1.5.1 The structure of *Leptospira interrogans* CMS (*Lin*CMS)

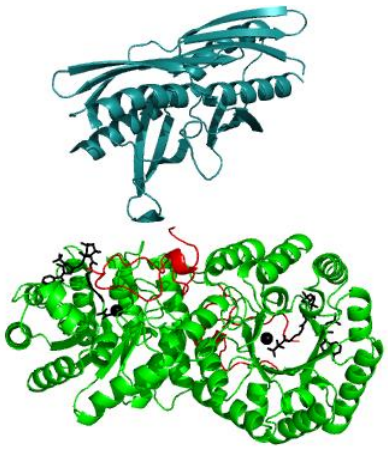


Figure 1.17: Two structures of *Lin*CMS (PDB: 3BLI, PDB: 3F6G).The catalytic domain is shown in green, a partial subdomain I is shown in red, and the regulatory domain (PDB: 36FG) is shown in teal. The substrates AcCoA and pyruvate, and the metal ion are shown in black.

*Lin*CMS is dimeric, and the monomeric structure is made of a catalytic $(\beta/\alpha)_8$-barrel barrel, two subdomains, and a C-terminal regulatory domain that is very similar in structure to the IPMS regulatory domain. Crystal structures of the catalytic domain with malonate, pyruvate, and with pyruvate and AcCoA have been solved (PDB codes 3BLE, 3BLF, 3BLI). The TIM barrel is very similar to that of *Mtu*IPMS and other IPMS structures solved, with a root mean square deviation (RMSD) of 1.8 Å between the *Lin*CMS and the *Mtu*IPMS barrels.[92] The substrate, pyruvate, is bound in a similar position to that of KIV in IPMS. A hydrogen bond is formed from pyruvate to a conserved threonine residue in the *Lin*CMS structures and to the metal ion that is coordinated to conserved histidine and aspartate residues, as in the *Mtu*IPMS crystal structures. Although the binding mode of the substrate is similar to that of KIV in the *Mtu*IPMS structure, the residues that comprise the active site differ between *Mtu*IPMS and *Lin*CMS to allow for control of substrate specificity. *Lin*CMS shows strong substrate specificity towards pyruvate, although it demonstrates some limited activity with other keto-acids such as ketobutyrate and ketoisovalerate.[102]

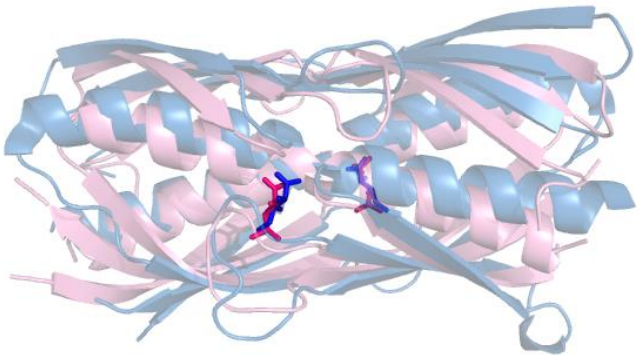


Figure 1.18: Allosteric ligand binding to the regulatory domain of *Mtu*IPMS .L-Leucine (blue) is bound in the interface of the regulatory domain of *Mtu*IPMS (PDB: 1SR9, light blue). L-Isoleucine (pink) is shown bound in the interface of the regulatory domain of *Lin*CMS (PDB: 3F6G, light pink)

CMS is allosterically regulated by L-isoleucine, which binds in the dimer interface of the regulatory domains as L-leucine does in IPMS (Figure 1.18).[78] It has been suggested that the mechanism of allosteric inhibition is similar between IPMS and CMS.[78] *Lin*CMS demonstrates K-type allostery towards both substrates, whereas *Mtu*IPMS demonstrates V-type allostery towards both substrates.[79, 103] However, there appears to be different mechanisms of allostery amongst the different IPMS proteins, suggesting that these enzymes may use a variety of mechanisms of allosteric regulation . The crystal structure of the regulatory domain of *Lin*CMS has been solved, showing L-isoleucine binding in a similar site to that of L-leucine in *Mtu*IPMS (Figure 1.18).

1.6 Homocitrate synthase

Homocitrate synthase (HCS) represents the first committed step in the AAA pathway, one of the pathways utilised by microorganisms to synthesise lysine.[71] Like both IPMS and CMS, the substrates for HCS are AcCoA and an ketoacid, in this case, ketoglutarate (Figure 1.19).[84]

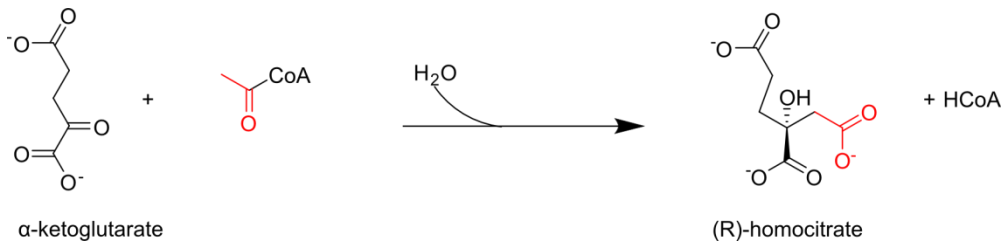


Figure 1.19: The reaction catalysed by homocitrate synthase.

HCS is structurally similar to both IPMS and CMS. Several full-length crystal structures of HCS have been solved from two organisms, the yeast *Schizosaccharyomyces pombe*, (*Spo*HCS) (PDBs 3IVT, 3IVS, 3IVU, and 3MI3) and a thermophilic bacterium, *Thermus thermophilus*, (*Tth*HCS) (PDBs 2ZAF, 3A9I, 2ZTJ, and 2ZTK).[104-106] These structures show that both *Spo*HCS and *Tth*HCS HCS have a very similar catalytic barrel to that of CMS and IPMS, and possess the corresponding subdomains attached to the barrel, although the structure of subdomain II could not be resolved in *Tth*HCS. Unlike the majority of IPMS enzymes, no HCS enzyme identified thus far has a canonical regulatory domain.[106] Instead, kinetic studies have shown that these HCS proteins are competitively inhibited by lysine, the end product of the pathway, binding at the active site.[107]

1.6.1 The structures of HCS from *Schizosaccharomyces pombe* and *Thermus thermophilus*

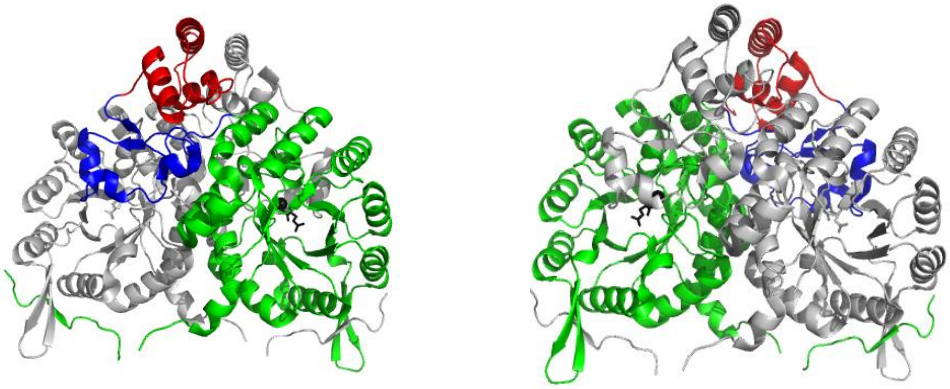


Figure 1.20: The structure of *Spo*HCS (PDB: 3IVT).The catalytic domain is shown in green, subdomain I in blue, and subdomain II in red. Chain A is shown in grey. The substrate, ketoglutarate, and the metal ion are shown in black. The right-hand image shows the enzyme rotated 180° compared to the left-hand image to highlight the contribution of chain A to the active site of chain B.

The structure of HCS from *Schizosaccharyomyces pombe* (*Spo*HCS) has been solved in the apo form, in two conformations with the natural substrate ketoglutarate (KG) bound in the active site (Figure 1.20), and with the competitive inhibitor, lysine, bound.[104, 105] The catalytic domain appears very similar to that of the *Mtu*IPMS and *Lin*CMS, although there is a short N-terminal extension that is disordered in the apo structure of *Spo*HCS, but forms a short β-strand and a 3_{10} helix in the ketoglutarate-bound structures. The subdomains are fully resolved in all four structures of *Spo*HCS. As observed in *Mtu*IPMS, subdomain I of *Spo*HCS crosses over to form part of the active site with the opposing chain, creating a domain-swapped homodimer. In one of the two KG-bound structures, the so-called 'closed lid' structure (PDB: 3IVT), two 3_{10} helices form the lid that prevents access to the active site. As with *Mtu*IPMS, subdomain II is comprised of a three-helix bundle.

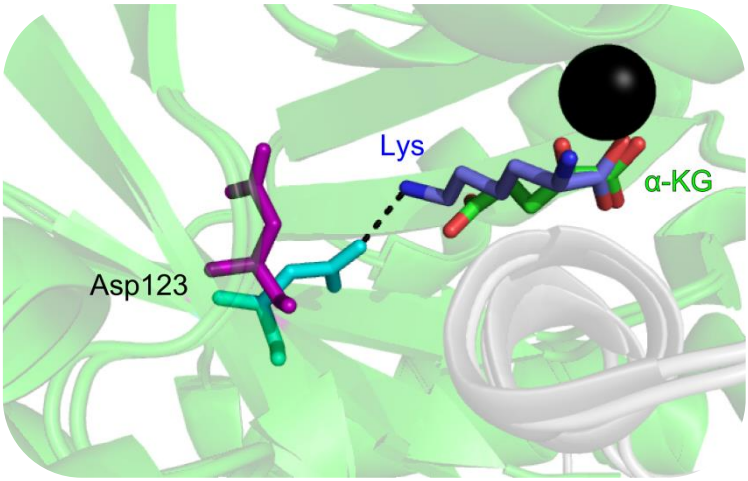


Figure 1.21: The movement of Asp123 in *Spo*HCS.Asp123 (cyan) when lysine (blue) is bound to the active site is compared the binding of ketoglutarate (Asp123 is shown in purple, KG in green).

In *Spo*HCS, there is a switch that allows acidic residues Asp123 and Glu222 to form interactions with the ε-ammonium group of lysine, the competitive inhibitor of HCS (Figure 1.21). In the ketoglutarate bound form, Glu222 facilitates the salt bridge formation of the side chain of Arg163 with the C5 carboxylate of ketoglutarate, and His103 additionally forms a hydrogen bond with the substrate. The aspartic acid is conserved in HCSs that lack a regulatory domain and thus are regulated by competitive inhibition.[69] Upon binding of lysine, there is also a slight conformational shift of the lid motif described above, where the lid motif moves away from the active site compared to the apo form, suggesting that the ligand bound in the active site can affect the conformation of the subdomains. Lysine is a competitive inhibitor towards ketoglutarate but a mixed inhibitor towards AcCoA, reaffirming the role of the subdomains in AcCoA interaction and binding[105].

Figure 1.22: The crystal structure of *Tth*HCS (PDB: 2ZYF).Chain A is shown in grey. The catalytic domain is shown in green, and subdomain I is shown in blue. The metal ion and ligand, KG, are shown in black. The right-hand image is a rotation of the left-hand image to highlight the contribution of the subdomains to the opposing chain's active site.

The crystal structure of *Tth*HCS has also been solved with ketoglutarate (Figure 1.22), lysine or the product, homocitrate, bound in the active site.[106] Unlike the crystal structures of *Spo*HCS, subdomain II in the *Tth*HCS structures was unable to be resolved, suggesting that it is very mobile. As with the previous structures, *Tth*HCS forms a domain-swapped homodimer, with the TIM barrel forming the catalytic domain, while subdomain I from the adjacent monomer contributes residues to the active site (Figure 1.22, right). The residues that contribute to ligand binding and specificity are similar to those in *Spo*HCS as described above, and as with *Spo*HCS, the binding of the inhibitor to the active site appears to rearrange the subdomains.

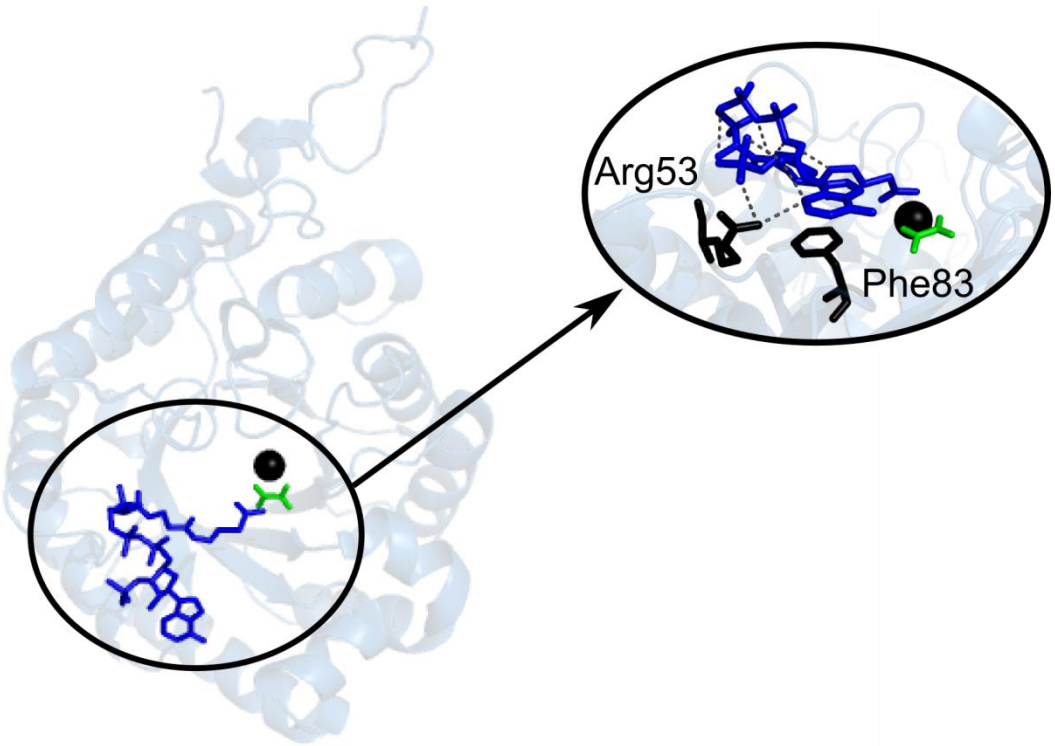


Figure 1.23: The structure of the catalytic domain of *Lin*CMS showing the binding of substrates.*Lin*CMS (light blue, cartoon, from PDB: 3BLI) showing AcCoA (blue), pyruvate (green), and the metal ion (black sphere). The zoomed circle shows interactions formed between AcCoA and Arg53 (black stick) and Phe83 (black stick).

A comparison was made between the pyruvate and AcCoA bound structures of *Lin*CMS and the ketoglutarate-bound structure of *Tth*HCS.[106] Although a large number of residues that are involved in AcCoA binding in the *Lin*CMS structure are conserved in *Tth*HCS, residues involved in recognition of the adenine ring and 3'-phosphoribosyl parts of AcCoA in *Lin*CMS are not conserved in *Tth*HCS, suggesting that there is a different binding mode for this large substrate in the different homologues. In *Lin*CMS, the adenine ring forms a π-π stacking interaction with Phe83 while the 3'-phosphate group of the adenine ribose forms an interaction with Arg53 (Figure 1.23).[102] By comparison, in *Tth*HCS, these residues are a leucine and a valine respectively.

1.6.2 Competitive inhibition in HCS

The regulatory domain of IPMS/CMS may be a comparatively new evolutionary addition, and the ancestral protein may have been unregulated or competitively inhibited as opposed to allosterically regulated. Due to the structural similarity between HCS and the allosterically regulated IPMS and CMS, the absence of a regulatory domain on characterised HCS enzymes suggests that IPMS and CMS may be able to perform catalysis without a regulatory domain, and indeed, this has been shown both in artificially truncated IPMS and CMS, and in the naturally truncated form.[2, 108] There is no alternative pathway for leucine biosynthesis, which suggests that an IPMS was the ancestral enzyme that was duplicated to allow for the evolution of CMS and HCS. As modern IPMS and CMS enzymes maintain the ability to catalyse reactions with the regulatory domain removed, it suggests that the ancestral protein did not maintain a regulatory domain. Drevland et al.[89] suggests there were two duplication events leading to the phylogenetically distinct types of CMS, and the difference in lysine inhibition and the great species difference between the *Thermus-Deinococcus* group of thermophilic bacteria and the fungi and yeast that utilise HCS and the α-aminoadipate pathway suggests there have also been multiple gene duplication events to produce two distinct types of HCS. Additionally, many organisms appear to maintain multiple copies of the *leuA* gene, with and without a regulatory domain, which suggests multiple gene duplication events.[2, 90]

Larson and Idnurm[90] demonstrated that the *leuA* gene from *Phycomyces blakesleeanus* was evolutionarily closer to those from plants and photosynthetic bacteria than it was to *leuA* genes from other fungi, although the gene product can complement a *leu3* (isopropylmalate synthase) *Schizosaccharyomyces pombe* knockout. As this species diverged early from other fungi during evolution, the gene in these organisms may be an ancestral variant while other organisms have acquired the gene more recently, or horizontal gene transfer events may have occurred, either to provide *Phycomyces blakesleeanus* with a *leuA* gene from a more modern cyanobacteria or plant, or to provide both lineages with different genes. This suggests another mechanism by which these genes, and pathways, can be transferred between organisms, and may suggest a mechanism by which diverse organisms such as thermophilic bacteria and fungi appear to have similar genes. These enzymes present an interesting picture of divergent and convergent evolution across both prokaryotes and eukaryotes.

1.7 Modular domain evolution

As more sequence and structural information becomes available, it has become apparent that non-homologous proteins can be built with domains common to multiple proteins, suggesting that the part of a gene that codes for a common domain can move through the genome as a modular unit.[109] There are numerous examples, particularly in eukaryotes where domains involved in protein-protein interactions and cell signalling appear in different proteins, yet fulfil the same role.[110] Additionally, there are examples where the same domain can perform different roles, for example, domains catalysing different chemical reactions, while retaining the same tertiary structure. The most ubiquitous example of this is the TIM barrel that makes up the catalytic barrel of the enzymes detailed above, and is the most common fold that has been structurally characterised.[111] The TIM barrel can catalyse a variety of different reactions, including the eponymous triose phosphate isomerisation of dihydroxyacetone phosphate and D-glyceraldehyde 3-phosphate.[70] Another example of a domain found in different contexts in a variety of proteins is the ACT domain.[112] Large scale domain rearrangements may have occurred in numerous different ways, from short duplications, to large, disruptive, chromosome mutations.[113]

1.7.1 The modularity of the IPMS and IPMS-like enzymes

Although the TIM barrel represents a common structural module in a variety of proteins and organisms, in the context of IPMS and IPMS-like enzymes, the catalytic module of these proteins contains the TIM barrel as well as subdomain I and subdomain II. Zhang et al.[2] argued that the catalytic and functional module of IPMS is the catalytic domain and the subdomains, due to the essentiality of the subdomains in facilitating catalysis. Interestingly, the fold that makes up the regulatory domain of IPMS has only been identified in IPMS and CMS to date. There is some evidence via databases such as Pfam[114] that it is also present in threonine synthase from *Francisella* species, but as of yet, these have not been characterised.[115] Moore et al.[113] suggests that these domains either emerged very recently, and thus have not spread through the genome, or have diverged so significantly in sequence that the current tools for identifying these domains are unable to identify them. It is interesting that this fold has been retained during the gene duplication events leading to the two phylogenetically distinct CMS enzymes, but was not in the HCS genes, although as discussed above, these genes may have been transferred through evolution by a variety of mechanisms that may account for this. Regardless, the idea of modular domain rearrangement

provides an interesting evolutionary tactic and a potential way to explore evolution of regulation in metabolic pathways.

Additionally, protein modularity presents an interesting way to investigate dynamics. In these proteins, the catalytic module includes two subdomains for which mobility is critical for catalysis, but these subdomains can bear the burden of a regulatory domain, and indeed, if this regulatory domain is removed, the proteins still facilitate catalysis.[2, 108] Therefore, studying how these proteins fluctuate in the presence and absence of the regulatory domain, both in the natural protein and in artificial constructs, may provide some clues as to how, during evolution, a highly dynamic protein has evolved to maintain a regulatory domain that restricts the conformations that the subdomains can form.

1.8 Summary

The study of protein conformation and dynamics is a field of increasing interest and importance. Investigating how proteins move is beginning to solve some of the mysteries that could not previously be solved with static techniques such as X-ray crystallography, and the study of dynamics is providing new insight into old conundrums such as allosteric regulation in the absence of a detectable conformational change. Understanding how proteins move, and how evolution mediates and shapes these movements, will be invaluable information for the design of antimicrobials, especially allosteric drugs.

IPMS, CMS, and HCS present an interesting evolutionary picture where dynamics plays a key role in both allosteric regulation and catalysis. The combination of allosterically and non-allosterically regulated proteins allows the exploration of how dynamics have evolved to mediate the burden of the regulatory domain while still facilitating catalysis, and how the dynamics of the subdomains can be altered in the presence of the allosteric inhibitor to attenuate catalysis at the distant active site.

A crystal structure of *Mtu*IPMS with L-leucine bound has been solved, and there is no significant conformational change when compared to the crystal structure without L-leucine bound. The lack of change suggests that a change in dynamics is of importance in transmitting the allosteric signal from the binding site in the C-terminal regulatory domain to the active site in the N-terminal catalytic domain. However, in the absence of crystallographic data, investigating the potential change in dynamics in phylogenetically diverse proteins such as *Nme*IPMS is difficult. A technique called statistical coupling analysis, that utilises sequence information, was used to investigate

whether a network of coevolved residues may contribute to the transmission of the allosteric signal in *Nme*IPMS-like IPMS proteins. The evolution of allosteric regulation was further explored by the construction of fusion proteins using the regulatory domains from *Lin*CMS and *Sso*HCS to investigate whether the allosteric network could be preserved even with a different regulatory domain.

Subdomain II of the IPMS and IPMS-like proteins, including HCS that lacks a regulatory domain, is critically important for catalysis, as removal of part of subdomain II renders the protein catalytically inactive. The structural similarity between the regulatory domain containing proteins and the proteins without a regulatory domain suggests that the overall mechanism of catalysis, of which subdomain II is a part, is similar. A previous truncation of *Nme*IPMS removed part of subdomain II as well as the regulatory domain and rendered the truncated protein inactive. To investigate whether a truncation that encompasses all of subdomain II can produce a truncated active *Nme*IPMS, the protein was truncated at the C-terminal end of subdomain II.

As extant IPMS enzymes have been found with and without a regulatory domain, it suggests that the dynamic subdomains can function under both structural constraints. However, the differences in dynamics between the two structural populations, especially in the subdomains, are of considerable interest, but structural information for both populations is lacking, making techniques such as molecular dynamics simulations difficult. Sequence information was also utilised to investigate coevolved residues in both populations to determine whether there may be a different network of coevolved residues that enable catalysis in both structural populations.

Therefore, in this study, the evolution and regulation of these three enzymes was investigated using a variety of techniques to explore how dynamics, and the allosteric control of dynamics, can evolve.

Chapter 2: Covariance analysis of IPMS

2.1 Introduction

Coevolution can be defined as two things affecting the evolution of each other. It is seen in a wide biological sense in things like the parasite-host interaction. Zaman et al.[116] discussed how the competition between parasite and host drives both parasite and host to evolve more complex functions, thus driving evolution of both species.

Coevolution, and covariation, can also be observed at the protein level. A change in residue or chemistry at one position may alter the local protein environment sufficiently to, upon random mutation at a second position; change the chemical environment acceptable in that second position.[117] Thus, a change in amino acid population at one position alters the amino acid population at another position. This is one way that antibiotic resistance can proliferate. Typically, the initial mutation selected for to confer resistance to the antibiotic can be harmful to the organism, causing a decrease in fitness.[117] However, the mutation can be maintained in the population if there is a compensatory mutation that may be deleterious by itself but is neutral or, less commonly, advantageous when combined with the first mutation, thus stabilising the resistance mutation in the bacterial population.[117] Therefore, the identification of coevolving networks in proteins is of importance to understanding antibiotic resistance and the development of new antimicrobial drugs.

Atchley et al.[118] describe potential sources of covariation in proteins. These include covariation with a phylogenetic basis, structural or functional covariation, and stochastic covariation. Stochastic covariation may include elements such as fluctuations in dynamics. Sfriso et al.[119] identified residue pairs that displayed coevolution but were not close structurally. They then utilised molecular dynamics and connected these residue pairs to other conformations of the protein, displaying that dynamics can also be a driver of coevolution.

Many different algorithms have been developed to detect coevolution in multiple sequence alignments. This project focuses on the most common methods: statistical coupling analysis (SCA) and mutual information (MI). MI will be discussed in depth in Chapter 3.

2.1.1 Introduction to statistical coupling analysis (SCA)

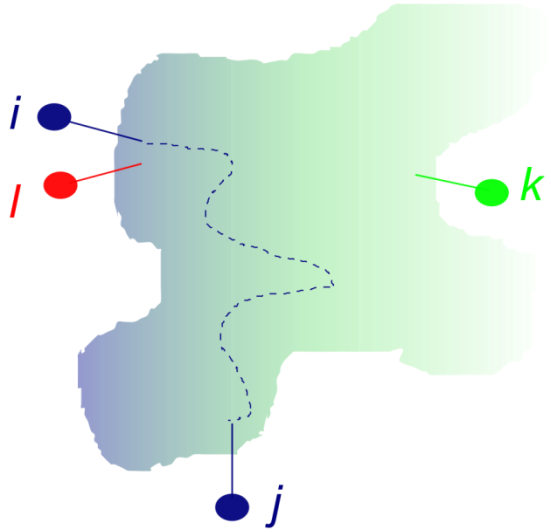


Figure 2.1: A theoretical protein demonstrating the difference between residues that show coevolution and those that do not.Position *l*, that is not involved in protein structure or function, position *k*, that is conserved but not coevolved with these residues, and positions *i* and *j*, that show coevolution

SCA allows the identification of networks of coevolved residues amongst evolutionary noise. An example is displayed in Figure 2.1, where a protein contains four generic residues named *i*, *j*, *k*, and *l*.[120] If the amino acid distribution at *l* is superfluous to both structure and function, then the amino acid frequency at this position should approximate the mean of all amino acid frequencies in all proteins. However, if residues, such as *i*, *j*, and *k* have a functional role, then the amino acid distribution at these positions should be different from the mean. This can be identified by simple conservation of residues at positions important for catalysis, for example, as represented by position *k*. However, if two residues are coevolved, they may not be absolutely conserved, but the distribution of residues at one site is affected by the distribution at another site. The degree to which one residue is affected by another can represented by a numerical score. The higher the score, the more coevolution shown by two residues. In this example, these residues are represented by *i* and *j*. Conversely, if the population of residues at position *k* is not affected by the population at either *i* or *j*, these residues have not coevolved even though position *k* shows significant conservation.

SCA was first pioneered by Ranganathan et al..[121] The technique was demonstrated using a multiple sequence alignment (MSA) of the PDZ domain family. SCA applied to this MSA identified a network of coupled residues that extended from the peptide ligand binding site through the domain to surface residues. The pathway identified using this technique was confirmed by thermodynamic mutant cycle analysis, as well as a binding energy assay. It was suggested that this

pathway represented a way by which the domain could alter its energetic connectivity in response to ligand binding.

SCA has also been used to determine which residues are necessary for maintaining functionality in a protein fold.[122] A SCA was performed on a MSA of WW domain family sequences, and through this, a group of coevolved residues was identified. Protein sequences were designed based on this SCA analysis, and the artificial proteins were then tested for activity. The structural and thermal properties of the artificial proteins were also explored. These artificial proteins were functionally indistinguishable from native sequences, suggesting that the function of the protein was fundamentally encoded in these coevolved residues, even though some were quite distant from the ligand binding site.

Investigating allosteric communication pathways has been a major focus of SCA, as allosteric pathways can be very difficult to discern from sequence or structure alone. One example is the allosteric communication pathway in haemoglobin.[120] A SCA was performed on a MSA of 800 globin family sequences. Following hierarchical clustering of scores determined by SCA, two overlapping groups of statistically coupled residues were identified. The first of these groups, once mapped onto the 3D structure of haemoglobin, included residues that interacted with heme and those that formed part of the tetramer interface. Based on experimental data, these regions were known to be important for the switch from the T to R state in response to O_2. This analysis demonstrates that SCA can be used to identify allosteric networks that have already been validated by experimental data.

Not only can SCA be utilised to explore allosteric pathways already known, but it can also be used to identify novel allosteric pathways. A MSA of cysteine peptidases was used in SCA to identify groups of coevolved residues, termed sectors.[123] These residues surrounded the active site, and extended through the protein as a network. A known ligand, chondroitin sulfate, was shown to form multiple contacts with sector residues. Putative allosteric binding cavities were identified utilising other software. These binding pockets also included sector residues, and of these, 'site 6' as it was termed, was shown to bind a novel allosteric ligand.

IPMS, as discussed above, is allosteric and the mechanism for allostery is not well understood but may involve the control of dynamics of the protein. It is, therefore, plausible that there is a network of residues that may not be absolutely conserved themselves that control the dynamics of the protein in response to L-leucine binding. SCA was used to try to identify this potential network.

2.2 Statistical coupling analysis of isopropylmalate synthases

2.2.1 Multiple sequence alignment construction

The initial sequence population and the final curated multiple sequence alignment are critical to the efficacy of covariance analyses such as SCA.[124] The population needs to be of substantial diversity so subpopulations, where there is substantial variation from the main population in terms of sequence at any particular site, do not dominate the alignment and thus skew the results of the SCA towards that subpopulation. Thus, positions in the MSA that do not show conservation should have an amino acid distribution approximately equal to the mean amino acid distribution for all proteins. This means that there has been sufficient evolution to allow for statistical deviation from the mean amino acid distribution at any position to be significant. In addition, the alignment needs to be of sufficient size that random removal of some members of the sequence population does not alter the amino acid distribution at sites where there is not conservation.

To obtain sequences for SCA, IPMS sequences were obtained using PSI-BLAST (Position-Specific Iterative Basic Local Alignment Search Tool) in an iterative fashion using the protein sequence of *Neisseria meningitidis* IPMS (*Nme*IPMS) as the input sequence.[101] The genus *Neisseria* was excluded from this search as otherwise the results were overwhelmed with results from this genus, and this would mean that the results were skewed towards this subpopulation of sequences. Following acquisition of sequences, the collection of sequences was run through the CD-HIT webserver to eliminate sequences that had a greater than 90% similarity to other sequences in the sequence set.[125]

2.2.2 Cluster Analysis of Sequences

CLANS (Cluster Analysis of Sequences) was then performed on the sequence pool, containing approximately 1500 protein sequences, and a single cluster was selected for further analysis (Figure 2.2).[126] CLANS uses an all-versus-all BLAST search to calculate pairwise attraction values. These attraction values are then used to produce a force-directed graph that allows visualisation of the sequence space. In the case of this sequence population, there is a single cluster, and the sequences cluster more tightly in the centre of the cluster. This is similar to the clustering seen when sequence similarity is calculated and visualised in 2.2.5.

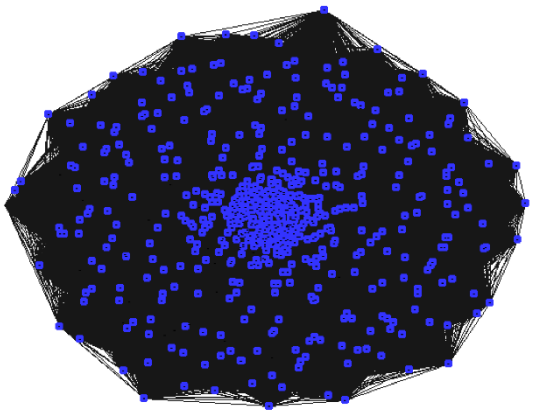


Figure 2.2: The graphical output of the CLANS analysis of the *Nme*IPMS-like IPMS sequence pool. This shows the clustering of the sequence population into a single cluster.

2.2.3 Multiple sequence alignment

The representative sequences of the CD-HIT and CLANS clusters, including *Nme*IPMS, were then used to perform an alignment using MAFFT.[92] MAFFT is a multiple sequence alignment programme that utilises fast Fourier transform (FFT) to perform alignments.[92, 127] The FFT-NS-2 algorithm was used for construction of the alignment; as of the algorithms available in MAFFT, this gave the best trade-off between relative speed and accuracy. In this heuristic method, a series of short 'words' of a defined length is identified in the sequence, and a distance matrix is computed based on the number of 'words' shared between any two pairs of sequences. The distance matrix is essential for the construction of a rough guide tree, from which an alignment can be constructed. In the FFT-NS-2 method, this alignment is then improved on by re-construction of the guide tree, followed by a second alignment that is typically more accurate than the first. MAFFT was selected as the programme used to construct the MSA as it performs well with different sequence populations.[128]

Following construction of the alignment, sequences that were less than 350 amino acids or more than 700 amino acids were removed. Sequences less than 350 amino acids typically were partial proteins, and as the typical length of an IPMS without a regulatory domain is 380 – 400 amino acids long, they were unlikely to be functional IPMS. Sequences of greater than 700 amino acids, of which there were very few, appeared to have additional domains or lacked a stop codon and were therefore not of interest for this study. Sequences that lacked the essential 'DRE' motif were also removed. This motif, consisting of Asp – Arg – Glu in a helix in the active site, has been shown to be essential for catalysis and is characteristic of the IPMS and IPMS-like proteins.[68].

Finally, sequences that lacked the regulatory domain were also removed as the interest in this particular analysis lies in proteins that display the canonical regulatory domain.

Iterative cycles of sequence removal and re-alignment to restrict the sequence alignment to proteins that are homologues of *Nme*IPMS was then performed. Those sequences with appropriately conserved residues at positions known to be important for substrate selectivity and inhibitor selectivity were maintained, based on residues identified by Hunter and Parker[72] and Kumar et al.[69] Active site residues, such as His204 and His206 that are critical for coordinating the metal ion, and Tyr313, known to be important for catalysis, show substantial conservation, while regions that do not appear to have a role in ligand binding or allostery do not show conservation. This gives confidence that the alignment accurately represents the pattern of conservation seen in IPMS sequences, including known regions of importance. A total of 584 sequences made up the final alignment, with a pairwise percentage identity of approximately 60%.

2.2.4 Removal of gaps and alignment with a known structure or model

The alignment was truncated in Matlab so any sequence position with a gap frequency of greater than 20% was removed. This prevented the trivial over-representation of gaps in the alignment in the final calculations. The sequence from the *Nme*IPMS PDB file was then aligned with the *Nme*IPMS sequence in the multiple sequence alignment. This produced a file that relates alignment numbering to structural numbering to allow for ease of later analysis.

2.2.5 Analysing sequence similarity and conservation

A matrix (S) was constructed comparing the similarity of pairs of sequences. Within the matrix, each position is the fraction of residues that are similar between two sequences in the multiple sequence alignment. The matrix can be represented in the form of histograms (Figure 2.3, left), or as a heat map (Figure 2.3, right). These suggest that the population is not completely homogenous with some obvious grouping. This is also reflected in the CLANS results, where there was a tight group of sequences in the centre, but the sequences clustered tightly together without additional clumping or clustering. These two results suggest that the sequence population can be used for SCA.

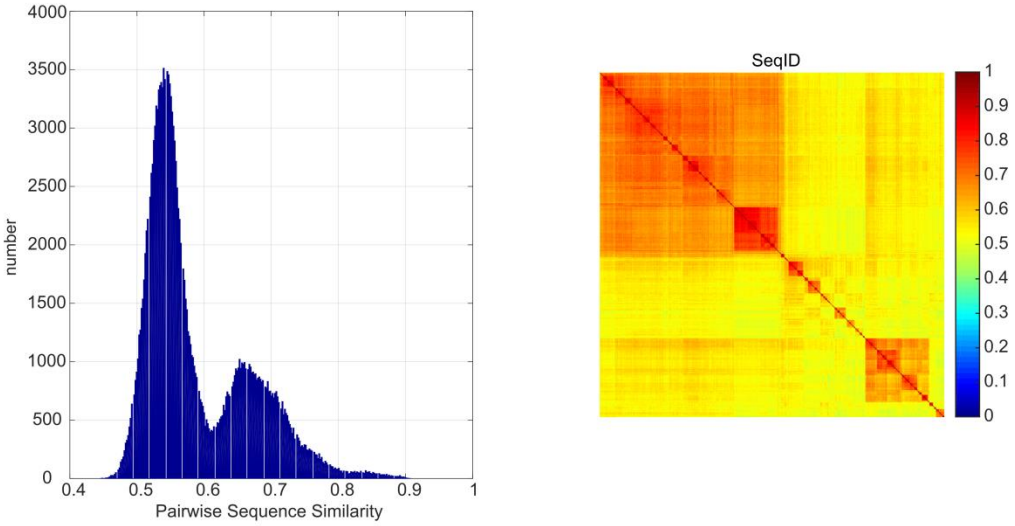


Figure 2.3: The similarity of sequences in the MSA used for the SCA.

The degree of conservation at each position was also assessed (Figure 2.4).[120] This was used primarily to assess how well the alignment matched with what is already known about the system, for example assessing whether conservation at key residues such as the histidines responsible for metal binding, His204 and His206, had been preserved in the alignment. The positional correlation was measured by a statistical quantity, D, which is the probability of observing the frequency of an amino acid at a position in an alignment of M number of sequences, given a specific background probability that has been computed from the mean frequency of all protein sequences in non-redundant databases.[120] D is also known as the Kullback-Leibler relative entropy.[120] The positional correlation scores were then mapped onto the homology model of *Nme*IPMS (Figure 2.4, bottom). The positional correlation scores matched well with previous alignments and what is known about the system. The conserved histidine residues at positions 204 and 206 show a high score for positional correlation, as do other conserved residues in the active site. Conversely, regions without substantial conservation, such as parts of subdomain II, have comparatively low positional correlation scores.

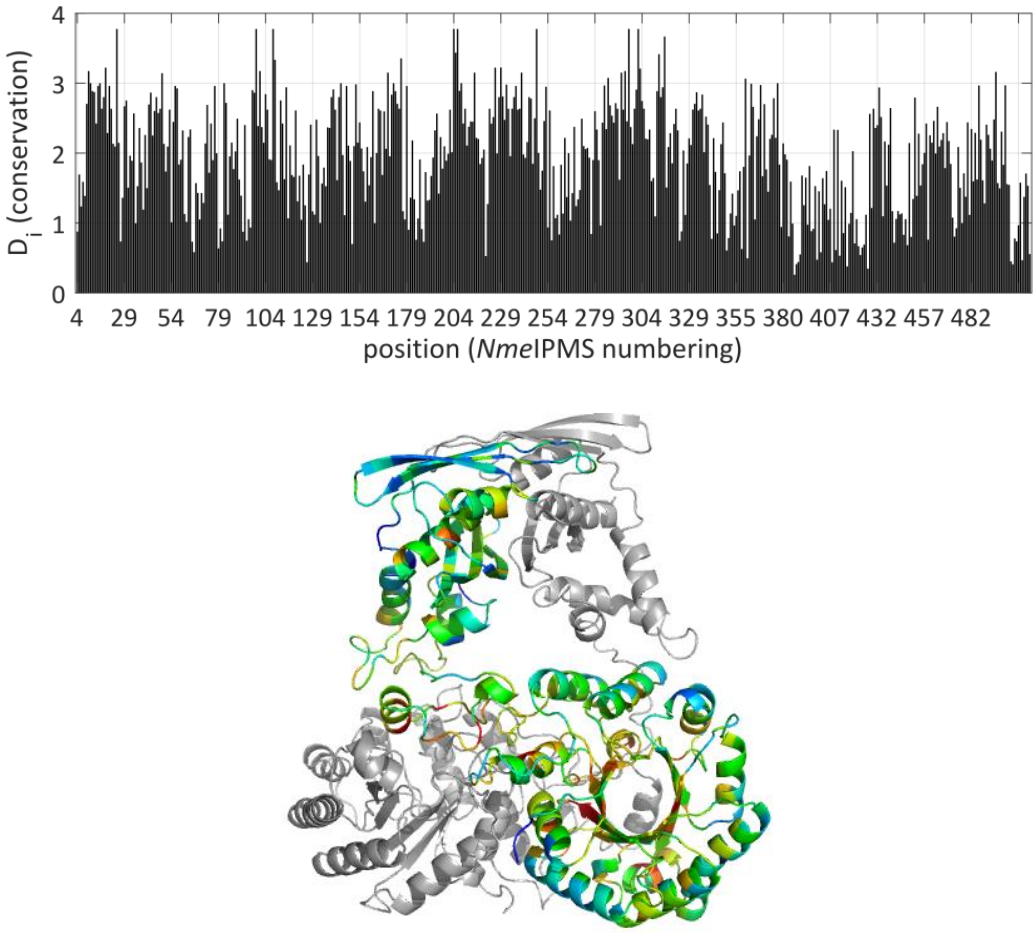


Figure 2.4: Positional correlation in the *Nme*IPMS-like IPMS used for the SCA.The positional correlation based on an entropy score is shown as a histogram (top) and mapped onto the homology model of *Nme*IPMS (bottom) where the highest entropy scores are shown in red and the lowest in blue.

2.2.6 SCA calculations

The SCA programme (sca5.m) was used to produce a positional correlation matrix (Cp), which numerically demonstrates the correlated evolution of all pairs of positions, and a sequence correlation matrix (Cs), which shows the pattern of similarity between all pairs of sequences. Unlike the sequence similarity matrix described in 2.2.5, these matrices are weighted based on conservation, so a change in a more conserved residue is weighted higher than one in a very variable position. The weighting provides the benefit of reducing noise. Residues that abut each other typically show coupling to maintain essential structural and functional features, such as correct protein fold and level of hydrophobicity or hydrophilicity in that region, and this low-level coupling can present as significant noise in the analysis. However, conservation weighting can bias the result by a phylogenetically related group. For example, if a particular group of sequences

within the MSA shows a change in residues involved in substrate binding that by nature are highly conserved, these residues will be statistically coupled based on this analysis, even though the grouping is tied more specifically to phylogeny. This phylogenetic bias can effectively drown out signals that are not related to phylogeny and limit the usefulness of the analysis. Although this technique was used with this caveat in place, a limited phylogenetic pool was selected for this analysis. This substantially limits the impact of conservation weighting, and thus provides more widely meaningful results.

2.2.7 Principal component analysis

Principal component analysis was used to determine which residues form 'sectors' in the protein. This involves the deconstruction of the data into eigenvalues and eigenvectors. An eigenvector is a direction and an eigenvalue provides information on how varied the data is in that direction. Every eigenvector has an eigenvalue. In this analysis, the eigenvector (otherwise known as an 'eigenmode') is a weighted combination of residues, whereas the eigenvalue of that eigenvector indicates how statistically important that eigenvector is.

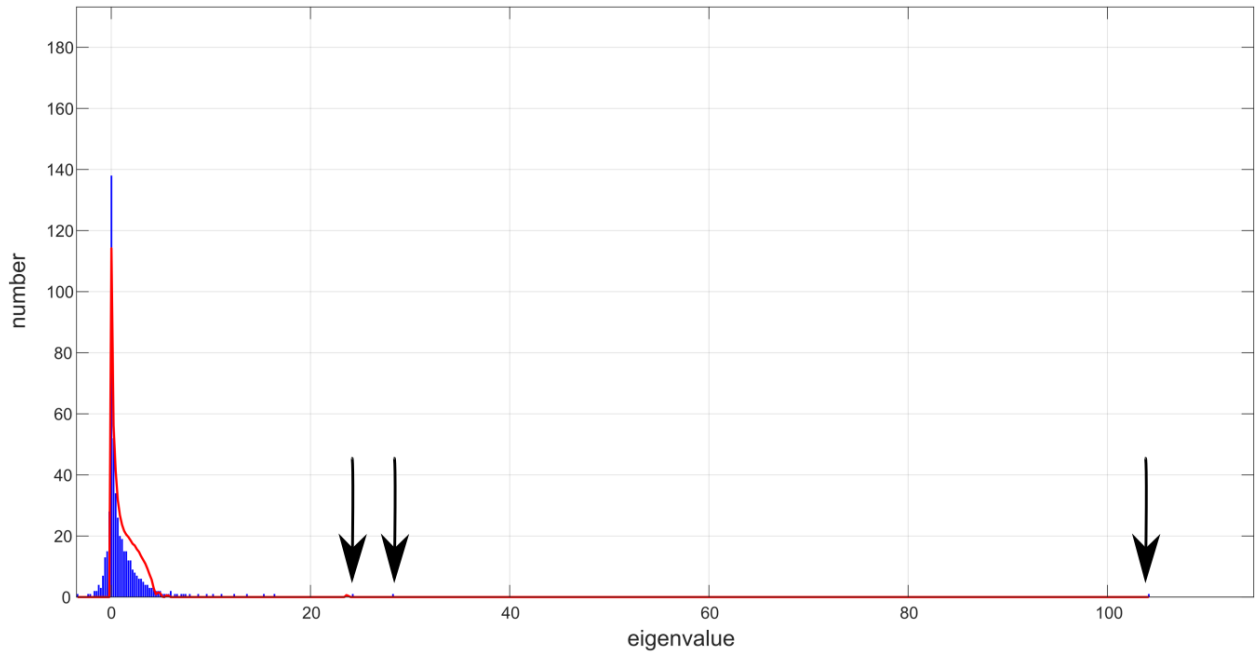


Figure 2.5: The eigenspectra of the principal component analysis of the *Nme*IPMS-like IPMS SCA. The arrows show the top three eigenmodes.

The eigenspectra, or distribution of eigenmodes by eigenvalue, was plotted (Figure 2.5). The alignments were also scrambled, with amino acids distributed at random in the columns of the MSA. This allowed 'randomised' alignments (red line, Figure 2.5) to be used to assess the degree

by which eigenvalues of eigenmodes could be explained by chance, as there is substantial stochasticity due to the nature of the data examined, and the number of sequences in the alignment is not significantly larger than the number of positions in the alignment. The top three eigenmodes are statistically significant, and the top eigenmode has a substantially larger eigenvalue than the others.

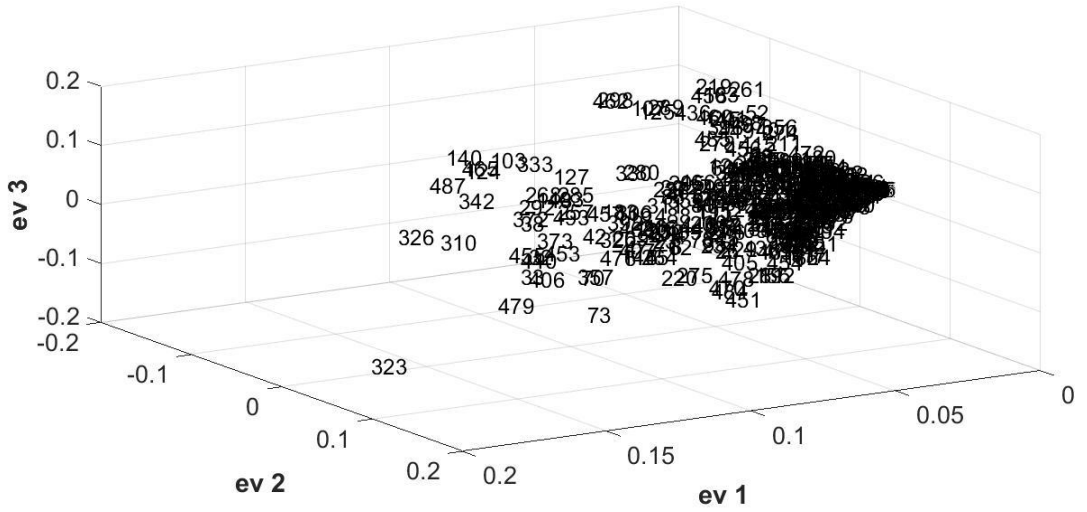


Figure 2.6: Scatter plot of the top three eigenvectors of the *Nme*IPMS-like IPMS SCA. The numbers represent the residue number from *Nme*IPMS.

The top 3 eigenmodes are represented as scatter plots in Figure 2.6, and show that most residues are not statistically coupled, and thus have small eigenvalues close to 0. However, some have comparatively large eigenvalues, and thus are residues of interest. To define the sectors, the eigenvalues in the top eigenmode were fitted to a lognormal distribution, and a cumulative density function was made from this (Figure 2.7). A cut-off was made at 70%, which meant the output represents residues in the top 30% of the cumulative density function, which form the tail. These were defined as the sector residues and are described below.

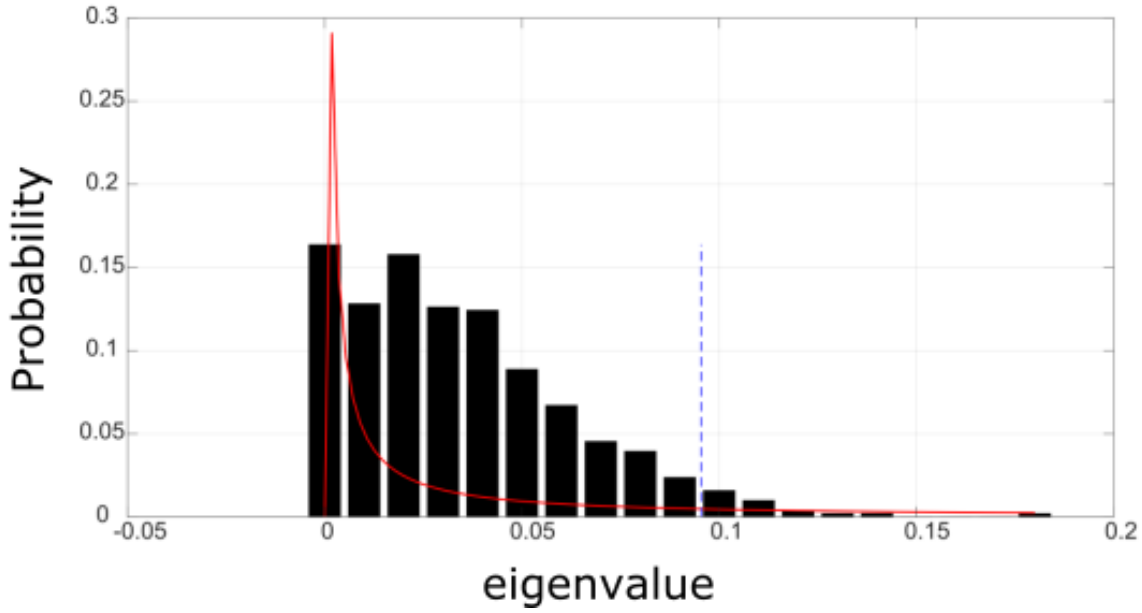


Figure 2.7: The eigenvalues of the top eigenmode of the *Nme*IPMS-like IPMS SCA(histogram) fitted to a lognormal distribution (red line). The blue dashed line shows the cut-off in the tail in the cumulative density function (CDF).

Additionally, positional correlation can be related to sequence correlation using single value decomposition to determine whether the sector residues were related to a phylogenetic group. Figure 2.8 shows that there is no particular definition of the sequence correlation, suggesting that there is only one relatively homogeneous sequence population. This suggests that the sector identified in Figure 2.7 is a global property of this sequence family, and not related to phylogenetic subsets. If there was a relationship between phylogeny and positional correlation, the results would look similar to those found in Halabi et al.[129], where several independent sectors were identified and these were directly related to phylogenetic relationships.

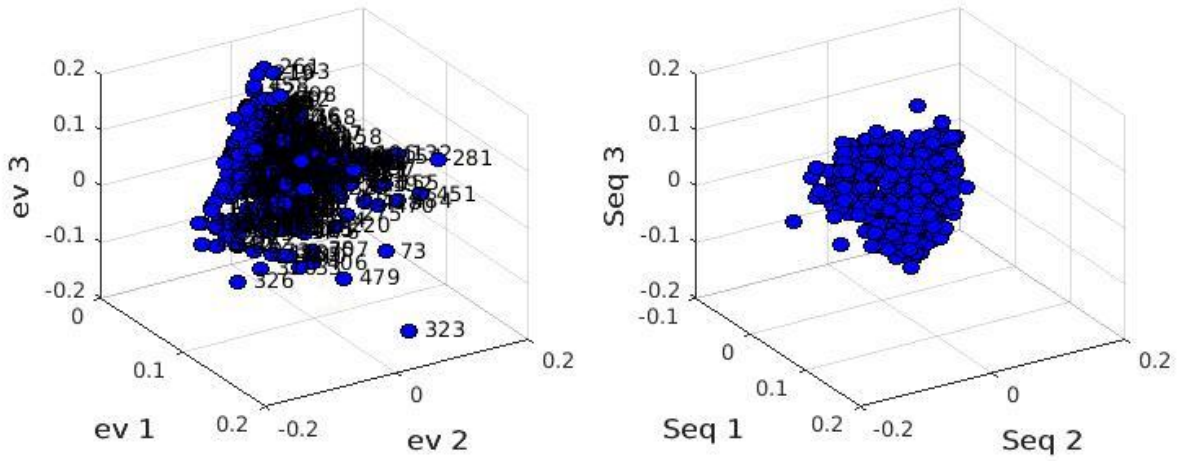


Figure 2.8: The top three eigenvectors (left) compared to the sequence correlation (right).

2.2.9 The structural and dynamic basis of the sector identified by SCA

A single sector, that does not appear to have a phylogenetic bias in this sequence population, was identified by principal component analysis. This sector was then mapped onto the homology model of *Nme*IPMS (Figure 2.9) and compared to results from molecular dynamics simulations that were performed by Dr. Wanting Jiao (personal communication, May 2014).

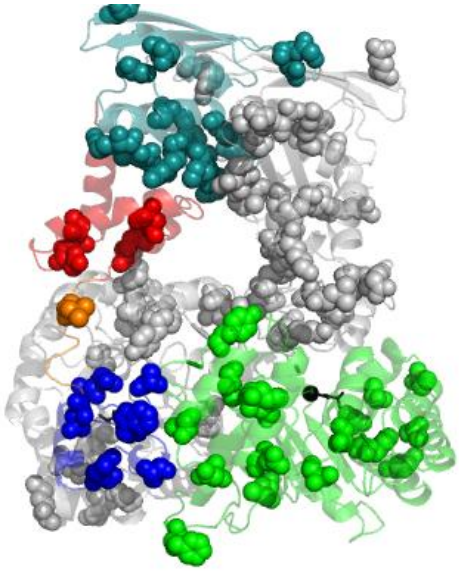


Figure 2.9: The single sector identified by PCA of the SCA mapped onto the *Nme*IPMS homology model.

The sector appears to contain residues spanning from the N-terminal catalytic barrel to the regulatory domain at the C-terminal end of the protein. Residues within the sector appear to form a network of interconnected residues that link the active site with the allosteric site in the interface between the two regulatory domains. The sector contains a large number of residues in the flexible subdomains, suggesting a potential mechanism by which the allosteric signal is transferred to the active site via the subdomains.

Table 2.1: Table of residues identified in the *Nme*IPMS SCA, as well as MD experiments performed on *Nme*IPMS (performed by Dr. Wanting Jiao), and the results of HDX experiments performed on *Mtu*IPMS by Frantom et al..[78]

Residue identified in *Nme*IPMS-like IPMS SCA	Residue location	Residue pairs identified in *Nme*IPMS MD simulations	Residue location	Residues identified in *Mtu*IPMS HDX experiments (*Mtu*IPMS numbering)	Location of residues identified in MtuIPMS HDX experiments
32	Catalytic domain	*Ligand free MD*		**78-87**	
33	Catalytic domain	Arg470(A)-Glu353(A)	Regulatory domain, subdomain II	79	Catalytic domain
38	Catalytic domain	Arg362(A)-Glu503(A)	Subdomain II, regulatory domain	80	Catalytic domain
70	Catalytic domain	Arg336(B)-Glu29(A)	Linker, catalytic domain	81	Catalytic domain
73	Catalytic domain	Arg32(A)-Asp375(B)	Catalytic domain, subdomain II	82	Catalytic domain
103	Catalytic domain	Ser20(B)-Glu298(A)	Catalytic domain, subdomain I	83	Catalytic domain
115	Catalytic domain	Lys342(B)-Glu29(A)	Subdomain II, catalytic domain	84	Catalytic domain
124	Catalytic domain	Tyr313(B)-Glu143(A)	Subdomain I, catalytic domain	85	Catalytic domain
140	Catalytic domain	Val454(A)-Val454(B)	Regulatory domain, regulatory domain	86	Catalytic domain
148	Catalytic domain	Glu298(A/B)-Glu236(B/A)	Subdomain I, catalytic domain	87	Catalytic domain
155	Catalytic domain	Ser299(A/B)-Glu236(B/A)	Subdomain I, catalytic domain	**453-457**	
220	Catalytic domain	Gln19(A/B)-Glu298(B/A)	Catalytic domain, subdomain I	454	Subdomain II
257	Catalytic domain	Ala456(A/B)-Tyr452(B/A)	Regulatory domain, regulatory domain	455	Subdomain II
264	Catalytic domain	Ser453(A/B)-Asn455(B/A)	Regulatory domain, regulatory domain	456	Subdomain II

268	Catalytic domain	Arg470(B)-Glu353(B)	Regulatory domain, subdomain II	457	Subdomain II
275	Catalytic domain	*Leu bound*		**488 – 495**	
281	Catalytic domain	Arg470(A)-Glu353(A)	Regulatory domain, subdomain II	489	Subdomain II
292	Subdomain I	Ser352(A)-Glu466(A)	Subdomain II, regulatory domain	490	Subdomain II
298	Subdomain I	Arg32(A)-Asp375(B)	Catalytic domain, subdomain II	491	Regulatory domain
310	Subdomain I	Arg371(B)-Glu58(A)	Subdomain II, catalytic domain	492	Regulatory domain
320	Subdomain I	Tyr281(B)-Pro282(A)	Catalytic domain, catalytic domain	493	Regulatory domain
323	Subdomain I	Lys332(A/B)-Glu18(B/A)	Linker, catalytic domain	494	Regulatory domain
326	Subdomain I	Val454(A/B)-Val454(B/A)	Regulatory domain, regulatory domain	495	Regulatory domain
333	Linker	Thr461(A/B)-Asp433(B/A)	Regulatory domain, regulatory domain	**617-632**	
342	Linker	Lys332(A/B)-Asp56(B/A)	Linker, catalytic domain	618	Regulatory domain
357	Subdomain II	Gly460(A/B)-Leu449(B/A)	Regulatory domain, regulatory domain	619	Regulatory domain
372	Subdomain II	Arg470(B)-Glu349(B)	Regulatory domain, subdomain II	620	Regulatory domain
373	Subdomain II	*Interactions broken in Leu bound MD compared to the apo system*		621	Regulatory domain
406	Regulatory domain	Arg362(A)-Glu503(A)	Subdomain I, regulatory domain	622	Regulatory domain
407	Regulatory domain	Arg336(B)-Glu29(A)	Linker, catalytic domain	623	Regulatory domain
421	Regulatory Domain	Gln19(A/B)-Glu298(B/A)	Catalytic domain, subdomain I	624	Regulatory domain
440	Regulatory domain	Glu298(A/B)-Glu236(B/A)	Subdomain I, catalytic domain	625	Regulatory domain

445	Regulatory domain	Ser20(B)-Glu298(A)	Catalytic domain, subdomain II	626	Regulatory domain
451	Regulatory domain	Tyr313(B)-Glu143(A)	Subdomain I, catalytic domain	627	
453	Regulatory domain	Arg470(B)-Glu353(B)	Regulatory domain, subdomain II	628	
455	Regulatory domain	*Interactions formed in the Leu bound MD compared to apo*		629	
462	Regulatory domain	Ser352(A)-Glu466(A)	Subdomain II, regulatory domain	630	
465	Regulatory domain	Lys332(A/B)-Glu18(B/A)	Linker, catalytic domain	631	
470	Regulatory domain	Thr461(A/B)-Asp433(B/A)	Regulatory domain, regulatory domain	632	
476	Regulatory domain	Arg470(B)-Glu349(B)	Regulatory domain, subdomain II		
478	Regulatory domain				
479	Regulatory domain				
484	Regulatory domain				
487	Regulatory domain				
493	Regulatory domain				

Hydrogen-deuterium exchange (HDX) studies in the presence and absence of L-leucine in *Mtu*IPMS identified several parts of the protein that undergo decreased exchange in the presence of L-leucine.[78] These regions include part of the regulatory domain surrounding the L-leucine α-amino group (residues 617 – 632, *Mtu*IPMS numbering), a group of residues in the regulatory domain that contact subdomain II (residues 488 – 495), a hydrophobic patch in subdomain II (residues 453 – 457), and conserved residues implicated in substrate and metal binding in the active site (residues 78 – 87) (Figure 2.10, Table 2.1). As the residues identified by HDX in the catalytic domain are conserved, they will not be identified by statistical coupling analysis, but if there is variation in the other residues, there could be statistical coupling identified in *Nme*IPMS using SCA if this is an allosteric path that is conserved between all IPMSs.

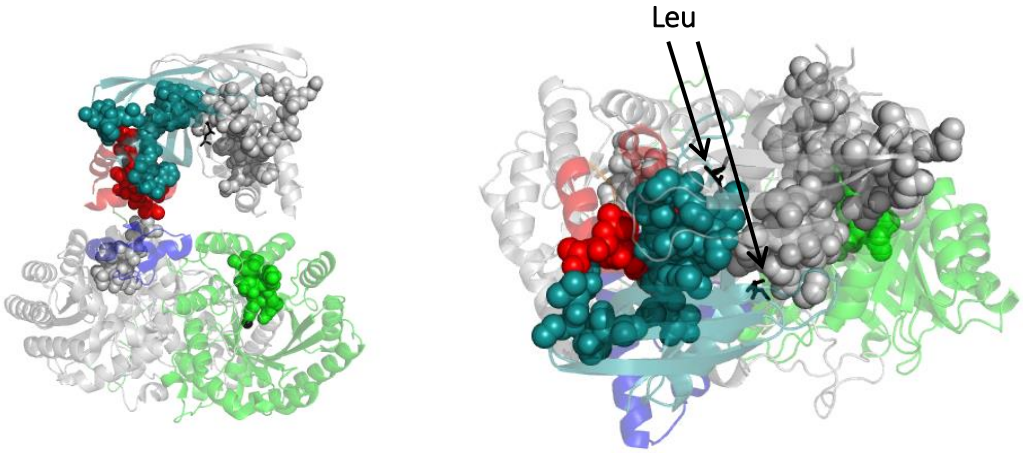


Figure 2.10: The structure of *Mtu*IPMS (PDB: 3FIG), residues identified by H/D exchange as showing a change in dynamics in the presence of L-leucine (spheres). Chain A is shown in grey, the catalytic domain is shown in green, subdomain I in blue, subdomain II in red, and the regulatory domain in teal. The Zn ion in the active site is shown as a black sphere. L-Leucine bound to the regulatory domain is shown as black sticks.

The residues 617 – 632 (*Mtu*IPMS numbering), identified by HDX of *Mtu*IPMS, that surround the α-amino group of L-leucine in the regulatory domain, and additionally form contacts with subdomain II, do show statistical coupling in the *Nme*IPMS-like IPMS SCA (residues 477 – 492, *Nme*IPMS numbering), but the other groups of residues identified by the hydrogen-deuterium exchange do not show statistical coupling. In the *Nme*IPMS-like IPMS alignment, the region spanned by residues 386 – 389 (*Nme*IPMS numbering), that corresponds to residues 488 – 495 in *Mtu*IPMS in subdomain II, has a considerable number of gaps, suggesting that this region is not highly conserved in length or sequence even in sequences closer in phylogeny than *Mtu*IPMS and *Nme*IPMS.

The region of subdomain II (residues 356 to 360, *Nme*IPMS numbering) of *Nme*IPMS that corresponds to residues 453 – 457 from *Mtu*IPMS shows a relative amount of conservation,

particularly at residue 360 (*Nme*IPMS numbering), but none of the residues show statistical coupling in the *Nme*IPMS SCA. This suggests that, if this set of statistically coupled residues identified in the *Nme*IPMS-like IPMS SCA forms an allosteric network, there is a difference in allosteric networks between *Mtu*IPMS and the *Nme*IPMS-like IPMS.

A difference in allosteric network between *Mtu*IPMS and *Nme*IPMS corresponds well with reports of different modes of allosteric inhibition in different branches of the phylogenetic tree of these enzymes. *Lbi*IPMS2, an *Nme*IPMS-like IPMS, shows V-type allostery towards KIV but a mixed V/K-type allostery towards AcCoA, while *Mtu*IPMS shows V-type allostery towards both substrates and slow-onset inhibition not seen in other IPMSs, while *Mja*IPMS shows V-type allostery towards both substrates and a similar mode of action to *Mtu*IPMS, that is altering the rate-determining hydrolytic step of the enzymatic reaction.[2, 78] The difference in the mode of allostery displayed by different populations of IPMS enzymes suggests that there may be a phylogenetic basis to the difference in allosteric networks and modes of regulation.

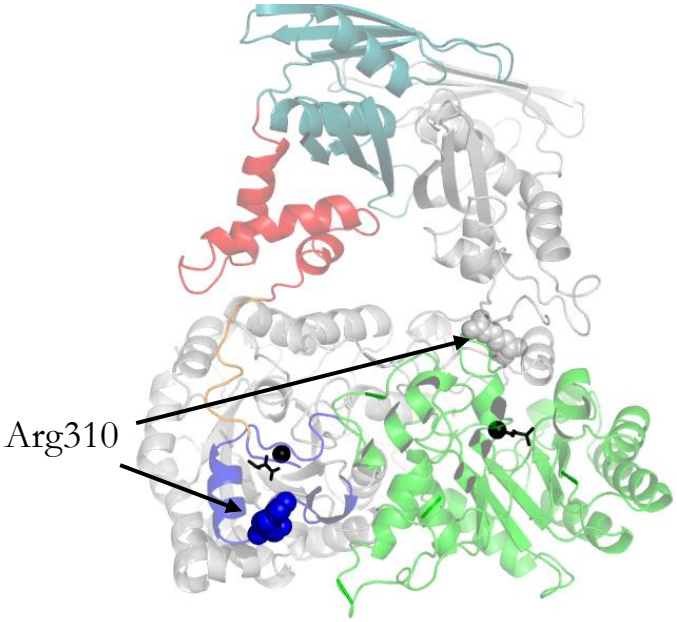


Figure 2.11: The *Nme*IPMS homology model with Arg310 highlighted as spheres.Chain A is shown in grey, Chain B is shown in colour: the catalaytic domain is shown in green, subdomain I in blue, the linker in orange, subdomain II in red, and the regulatory domain in teal. The active site is highlighted by KIV (black spheres) and the metal ions (black spheres).

Another residue identified in the sector identified by the *Nme*IPMS SCA is Arg310 which is located in subdomain I (Figure 2.11). An Arg310Ala mutation has been made in *Nme*IPMS, and this mutant shows a comparable response to L-leucine as that observed in wild type *Nme*IPMS, but Arg310Ala had a significantly higher K_m for AcCoA of 750 μM ± 260 compared to the wild-type protein, suggesting that Arg310 may not be part of the allosteric pathway but contributes

substantially to ligand binding.[130] However, substrate binding, especially AcCoA interaction and binding, are intimately linked with both allosteric regulation and movement of the subdomains so this residue may be statistically coupled to those involved in allosteric signal transmission without being involved itself.

To investigate the nature of this statistical coupled sector further, additional alanine mutations were made in *Nme*IPMS. These residues were chosen as they were identified in the MD simulations as forming potentially interesting interactions. Additionally, they also appear in the SCA, suggesting that the MD and SCA may have identified parts of similar networks of interaction that contribute to allostery.

2.3 Mutants in *Nme*IPMS based on MD and SCA

Several residues identified by the SCA of *Nme*IPMS-like IPMSs are particularly interesting. MD simulations conducted and analysed by Dr. Wanting Jiao (personal communication, May 2014) on the apo *Nme*IPMS homology model, and the model with L-leucine bound, were performed, and changes in hydrogen-bonding networks in the presence of L-leucine were identified. The only overlap between residues identified from the SCA and residues identified from the MD analysis are Arg470, Arg32, Glu298, Ser453 and Asn455 (Figure 2.12).

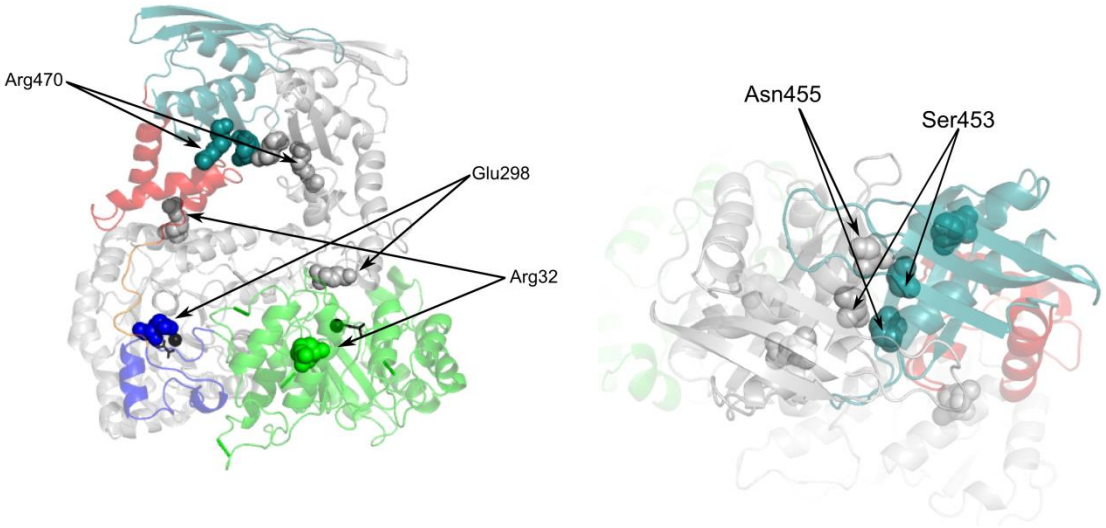


Figure 2.12: Residues identified in the SCA and in the MD simulation shown on the homology model of *Nme*IPMS. Chain A is shown in grey. The catalytic domain of chain B is shown in green, subdomain I in blue, the linker in orange, subdomain II in red, and the regulatory domain in teal. The residues of interest are shown as spheres. The active site is denoted by KIV (black stick), and the metal ion (black sphere).

In the MD simulations, both with and without L-leucine, Ser453 and Asn455 form a cross-chain interaction in the regulatory domain. When MD simulation was performed, potentially interesting residues, some of which showed a change in interaction in the apo simulation compared to the L-leucine bound simulation, were highlighted. In the MD simulations, Arg470 forms an interaction with either Glu353 or Glu349 in both the presence and absence of L-leucine, but is the only positive charge on the bottom of the regulatory domain, suggesting a potential role in allosteric regulation (Figure 2.14). Arg32 from chain A, in both the ligand free and L-leucine bound simulation, forms an interaction with Asp375 from chain B, making a potentially important link between subdomain II and the catalytic domain. Previous work has shown that mutation of Glu353 or Asp375, neither of which were identified in the SCA, to alanine only had a mild or no effect on L-leucine sensitivity.[130, 131]

The lack of change in response to L-leucine seen in the Glu353 and Asp375 mutants suggests that there are multiple factors in play when assessing the validity of both the MD simulations and the SCA analysis.

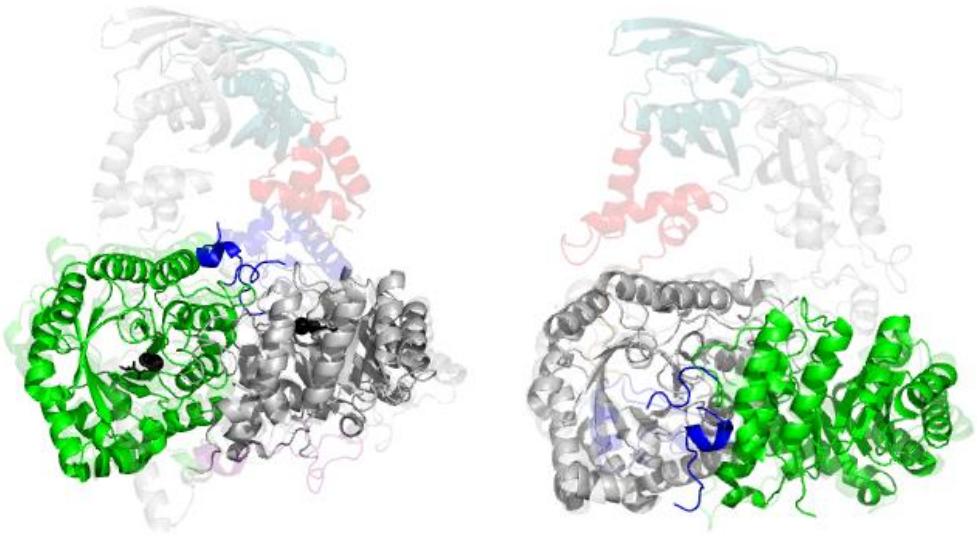


Figure 2.13: The structural alignment of the partial *Nme*IPMS crystal structure (PDB: 3RMJ) and a *Mtu*IPMS crystal structure (PDB: 1SR9)(left, transparent)and the *Nme*IPMS homology model(right, transparent).The catalytic domain is in green, subdomain I in blue, the linker in orange, subdomain II in red, and the regulatory domain in teal. Chain A is shown in grey. KIV (black stick) and the metal ion (black sphere) are shown in the left hand figure and denote the active site.

The MD simulations are limited as there are no structural data for the full-length *Nme*IPMS. The homology model used as the starting point for the MD simulation was constructed based on the *Mtu*IPMS crystal structure. The catalytic domain of the homology model, however, does structurally align with the partial crystal structure of *Nme*IPMS that has been solved (Figure 2.13) (RMSD: 1.96 Å). The partial crystal structure of *Nme*IPMS also aligns structurally with the *Mtu*IPMS crystal structure (Figure 2.13, RMSD: 1.51 Å). Additionally, the crystal structures of *Mtu*IPMS are the only full-length structures of IPMS, and, as discussed above, *Mtu*IPMS is evolutionarily distant from *Nme*IPMS, meaning that conformations that *Nme*IPMS adopts may not be observed in *Mtu*IPMS and vice versa.

The limitations of crystallography are also a problem as the conformation *Mtu*IPMS adopts in the crystal structure, which the *Nme*IPMS homology model is based off, may not be catalytically or allosterically relevant. Small-angle X-ray data suggests that *Mtu*IPMS adopts multiple conformations in solution[74]. This may mean that the initial conformation used for the *Nme*IPMS

MD simulations may not reflect the conformations adopted by the protein during the catalytic cycle.

There are also limitations of the SCA. SCA detects the change in distribution of amino acids at one residue position in response to the change in distribution of amino acids at another position. If a position is highly, or absolutely, conserved, even if it is biologically relevant and involved in the network, it may not show statistical coupling. Asp375, for example, is almost absolutely conserved in this alignment, so is unlikely to show statistical coupling. Additionally, SCA looks at the broader protein group to identify potential networks, but *Nme*IPMS may have undergone additional evolutionary pressures that affected the network, meaning the residues identified in the network may not entirely overlap with residues identified in the MD simulation that looks at only *Nme*IPMS.

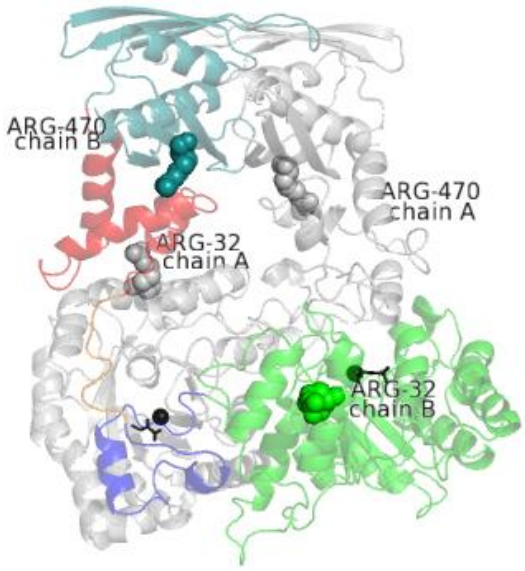


Figure 2.14: The homology model of *Nme*IPMS showing the locations of Arg32 and Arg470(spheres). Chain A is shown in grey, Chain B is shown in green (catalytic domain), blue (subdomain I), red (subdomain II), and teal (regulatory domain). The active site is denoted by the black spheres (metal ions) and the black sticks (KIV).

2.3.1 *Nme*IPMS Arg470Ala and Arg32Ala

Alanine mutants have been made for Arg470 and Arg32 in *Nme*IPMS.[132] The two mutants had been partially characterised previously.[131] Both Arg470Ala and Arg32Ala display similar Michaelis-Menten kinetics to the wild type protein (Table 2.2), although Arg470Ala showed a slight increase in K_m for both AcCoA and α-KIV compared to the wild type protein. Both mutants showed an approximately 2-fold decrease in k_{cat} compared to the wild type protein. Both mutant proteins

were insensitive to L-leucine inhibition up to 10 mM, suggesting that the connection between the regulatory domain and the catalytic domain had been severed, or that L-leucine was no longer binding the regulatory domain.

Table 2.2: Kinetic parameters for *Nme*IPMS wild type, *Nme*IPMS Arg32Ala, and *Nme*IPMS Arg470Ala. *Kinetic characterisation of the arginine mutants was performed by Matthew Plowman-Holmes

Enzyme	K_m (KIV, μM)	K_m (AcCoA, μM)	k_{cat} (s^{-1})
*Nme*IPMS wild-type	36 ± 3	35 ± 3	7.2 ± 0.1
*Nme*IPMS Arg32Ala*	44 ± 4	39 ± 4	3.1 ± 0.1
*Nme*IPMS Arg470Ala*	58 ± 6	55 ± 3	2.8 ± 0.1

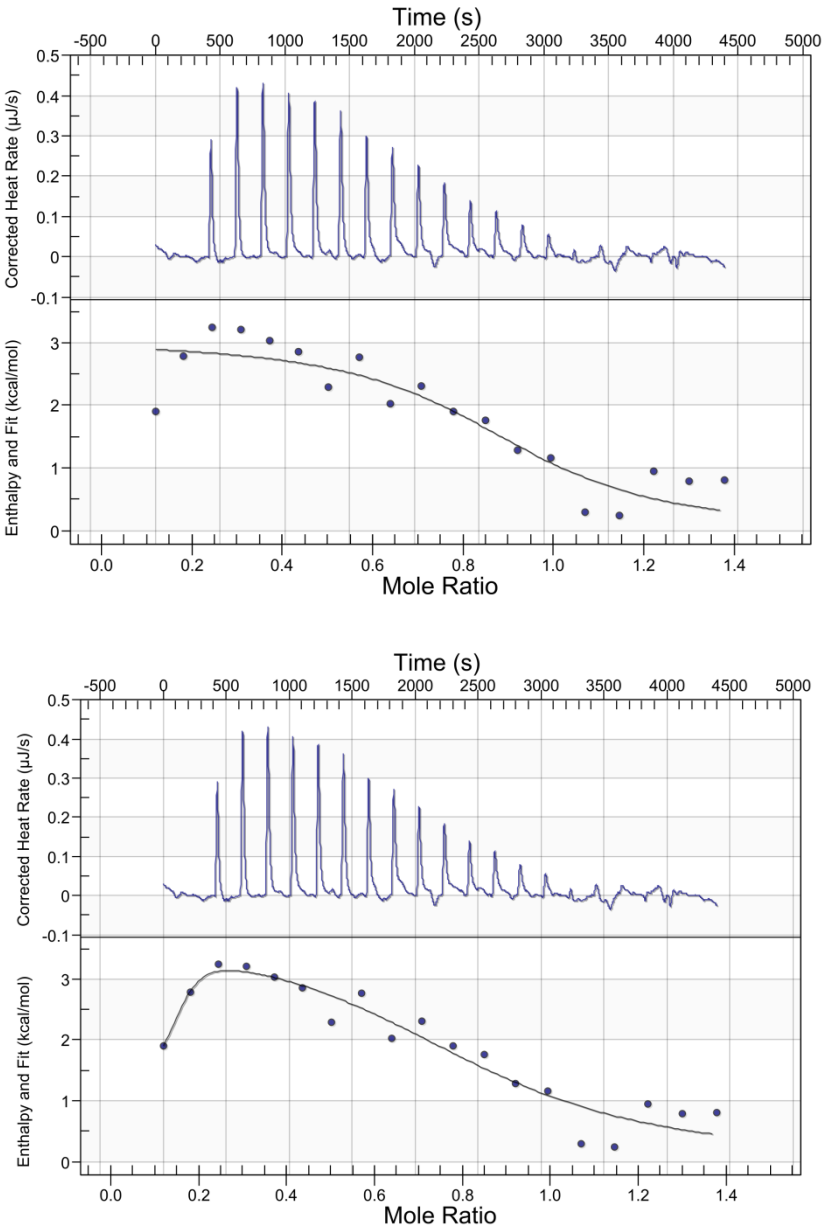


Figure 2.15: ITC data for *Nme*IPMS (100 µM) using L-leucine (400 µM) as the ligand.The data are fitted with an independent model (top) and a multiple site model (bottom). The thermodynamic parameters determined from these models are shown in Table 2.3.

Table 2.3: Thermodynamic parameters of the independent model and multiple site model fitted to the *Nme*IPMS wild-type ITC data

Model type	K_d (M)	n	ΔH (kJ/mol)	ΔS (J/mol.K)
Independent	$2.06 \times 10^{-6} \pm 3.21 \times 10^{-6}$	0.845 ± 0.09	12 ± 2	149

Model type	*Kd*1 (M)	*Kd*2 (M)	n1	n2	ΔH1 (kJ/mol)	ΔH2 (kJ/mol)	ΔS1 (J/mol·K)	ΔS2 (cal/mol·K)
Multiple sites	1.08×10^{-9}	2.2×10^{-6}	0.1 ± 0.6	0.6 ± 7	6 ± 90	11 ± 120	1.93×10^{2}	1.47×10^{2}

2.3.1.1 Isothermal titration calorimetry of Arg470Ala and Arg32Ala

Isothermal titration calorimetry was used to assess whether leucine was still binding the protein. Isothermal titration calorimetry (ITC) is a technique used to assess the binding of a ligand, in this case L-leucine, to a protein by measurement of the heat produced or absorbed by the protein upon ligand binding.[133]

A model could not be reliably fitted to the ITC data obtained for the binding of leucine to *Nme*IPMS, although an independent model and a multiple site model were both fitted to the data using NanoAnalyze (Figure 2.15, Table 2.3).

As the binding of leucine by *Nme*IPMS is endothermic, where ΔH is unfavourable but ΔS is favourable, heat is absorbed by the protein upon ligand binding. The increase in entropy implies an increase in flexibility in all or part of the protein, as well as a potential partial desolvation effect when leucine interacts with the protein. The binding curve of leucine binding to wild-type *Nme*IPMS matches well with previously reported ITC data of leucine binding to *Mtu*IPMS.[63] This similarity suggests that, although the absolute residues involved in the allosteric pathway are different, there may be a similar thermodynamic mechanism of allostery.[63] Interestingly, the Tyr410Phe *Mtu*IPMS mutant that is not inhibited by leucine, displays a similar binding curve to that of the wild-type protein, although there is a considerable increase in enthalpy upon leucine binding to the mutant compared to the wild type protein.

Popovych et al.[41] studied the binding of cyclic adenosine monophosphate (cAMP) to catabolite activator protein (CAP) using ITC. Although binding of cAMP to the dimeric CAP is negatively cooperative, binding of the first molecule of cAMP does not adversely affect the conformation of the second subunit but instead alters the dynamics of the protein, creating the negative cooperativity. Although cooperativity has not been suggested in the binding of leucine to *Nme*IPMS, the binding curve suggests that, in contrast to the canonical sigmoidal binding curve, other factors such as changes in dynamics are playing a role in the heat of "ligand" binding.

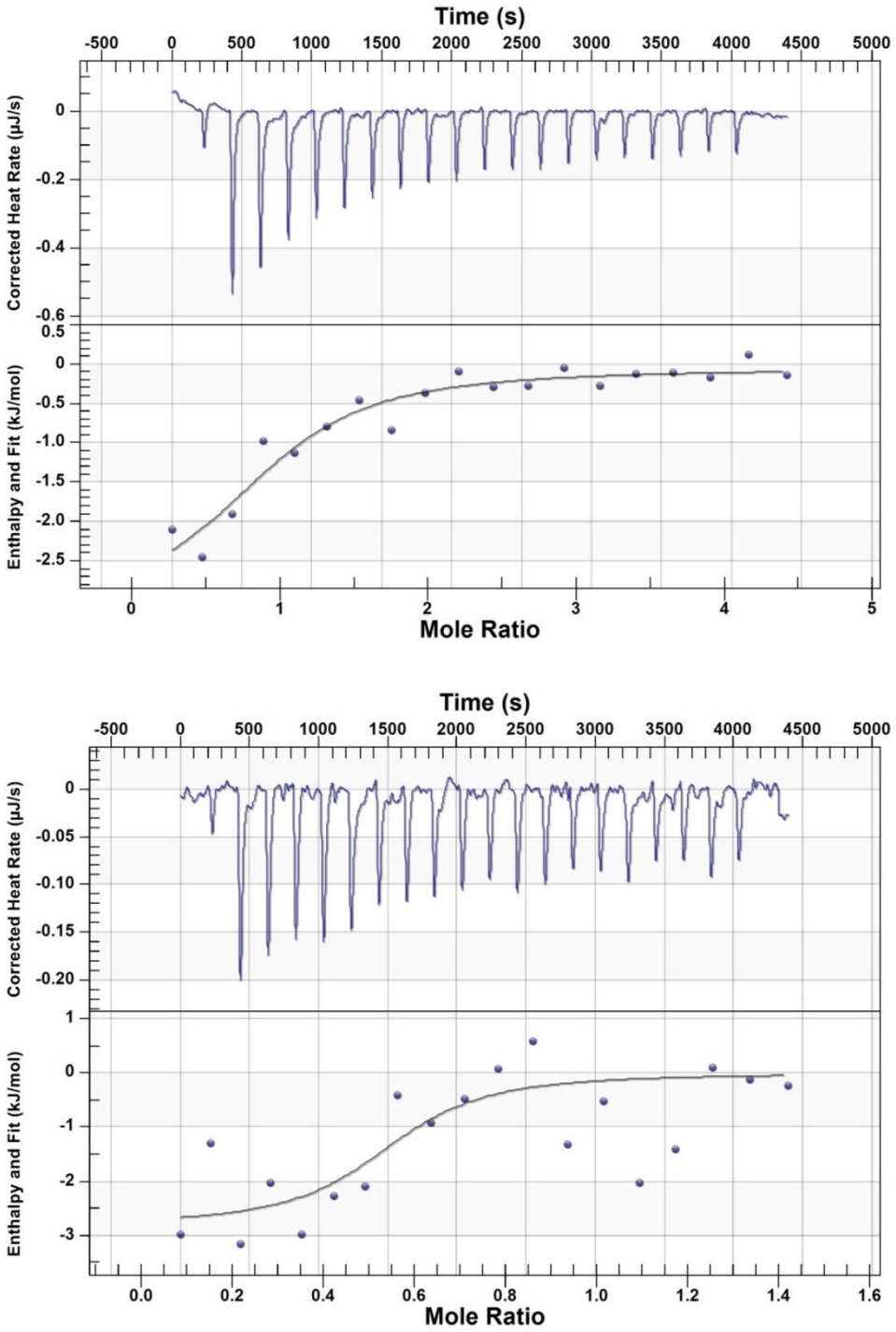


Figure 2.16: Isotherms of *Nme*IPMS Arg32Ala (top) and *Nme*IPMS Arg470Ala (bottom). A protein concentration of 150 µM and a ligand concentration of 2 mM was used for the Arg32Ala titration, and a protein concentration of 140 µM with a leucine concentration of 600 µM was used for the Arg470Ala titration.

Table 2.4: Thermodynamic parameters and stoichiometry of the *Nme*IPMS leucine insensitive mutants determined by ITC.

Mutant	Model	Stoichiometry	K_d (M)	ΔH (kJ/mol)	ΔS (J/mol.K)
*Nme*IPMS Arg32Ala	Independent	0.9 ± 0.8	$3.148 \times 10^{-5} \pm 1 \times 10^{-4}$	-3 ± 50	75.90
*Nme*IPMS Arg470Ala	Independent	0.5 ± 0.3	$3.145 \times 10^{-6} \pm 2 \times 10^{-4}$	-3 ± 20	165

Isothermal titration calorimetry was also performed for the two mutants that were not inhibited by leucine (Figure 2.16, Table 2.4). These mutants appear to bind leucine, although the binding curves are very different to that of wild type *Nme*IPMS. The ITC binding curve for both Arg470Ala and Arg32Ala is weakly exothermic as opposed to the endothermic binding curve of the *Nme*IPMS wild-type protein, demonstrated by the change from positive enthalpy in the binding of leucine to wild-type *Nme*IPMS to negative enthalpy in the binding leucine to both the mutants. Both mutants display much lower heats of binding than the wild type protein. As some form of binding is observed, it suggests that the two mutant proteins appear to be binding leucine, albeit, potentially, in a different way to the way *Nme*IPMS wild-type binds leucine. Other factors, such as changes in dynamics upon leucine binding, may also have been disrupted in the two mutant proteins.

2.3.2 *Nme*IPMS Glu298Ala

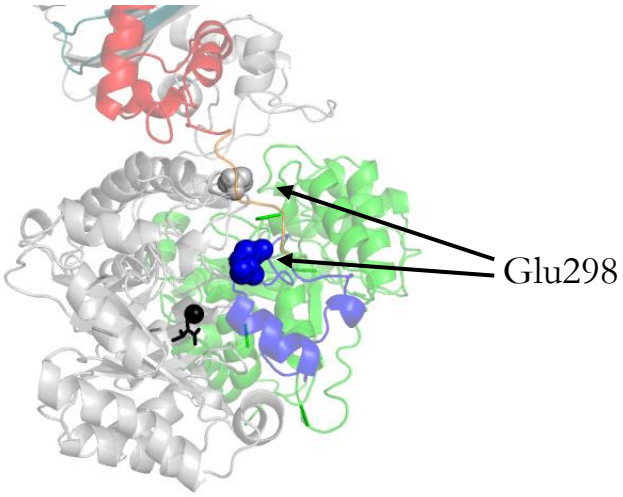


Figure 2.17: The location of Glu298 in the *Nme*IPMS homology mode. The residue is shown as spheres, and in Chain B (blue sphere) the active site is shown by KIV (black sticks) and the metal ion (black sphere), demonstrating the proximity of Glu298 to the active site

Glu298 is another residue of interest (Figure 2.17). As with Arg470 and Arg32, Glu298 was identified in both the SCA and in the MD simulations when performed both with and without leucine. In the ligand free MD simulation, Glu298 from chain A was predicted to form interactions with Ser20 of chain B and Gln19 in both monomers. As with the potential interaction partners of Arg470 and Arg32, the residues predicted to interact with Glu298 do not appear in the SCA. Both Gln19 and Ser20 are absolutely or highly conserved in the alignment (Figure 2.18). The residues may have additional roles aside from allostery, and the level of conservation may reflect this.

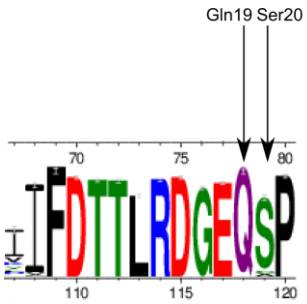


Figure 2.18: A LOGO diagram of the MSA used for the SCA showing the conservation of Gln19 and Ser20.

The equivalent to Gln19 in *Mtu*IPMS, Gln84, demonstrates a change upon L-leucine binding in one chain of the protein (Figure 2.19). In the allosteric ligand-free structure of *Mtu*IPMS, Gln84 forms an interaction with Arg427, and this interaction is broken in the leucine-bound structure, although the change in contact has not been experimentally linked to allostery. A Gln84Ala mutant

was made in *Mtu*IPMS, and this caused a large increase in the K_m for AcCoA and a substantial decrease in k_{cat}, but a more moderate impact on the K_m for KIV[68]. This mutation did not cause a significant impact on inhibition by leucine, suggesting that in *Mtu*IPMS, Gln84 is not involved in allosteric regulation by leucine.

The *Nme*IPMS Glu298Ala mutant was made to assess the impact of removing this charge on the kinetics of the enzyme but more specifically on the inhibition by leucine.

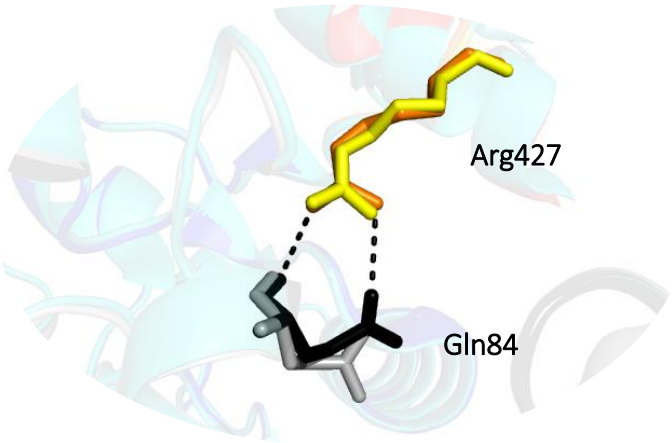


Figure 2.19: The interaction between Gln84 and Arg427 in *Mtu*IPMS in the absence of leucine(orange - Arg427, black - Gln84, PDB: 1sr9), and the presence of leucine (yellow - Arg427, grey – Gln84, PDB: 3fig)

2.3.2.1 Production of *Nme*IPMS Glu298Ala by site-directed mutagenesis

Whole circle site-directed mutagenesis was used to create the Glu298Ala mutant. The template used was the wild type *Nme*IPMS that had previously been cloned into pET151 via TOPO cloning.[64] Therefore, the expressed protein contained a N-terminal His_6 tag and a TEV protease site. The protein was expressed as soluble protein and was purified by IMAC followed by size-exclusion chromatography. The N-terminal His_6 tag was not removed, as discussed in Chapter 4.

2.3.2.2 Kinetic analysis of *Nme*IPMS Glu298Ala

The kinetic parameters of *Nme*IPMS Glu298Ala were determined using a chemically coupled assay, where the chemical couple, 4'-4'-dithiopyridine (DTP), interacts with the free thiol of CoA, the product of the reaction, and causes an increase in absorbance at 324 nm.

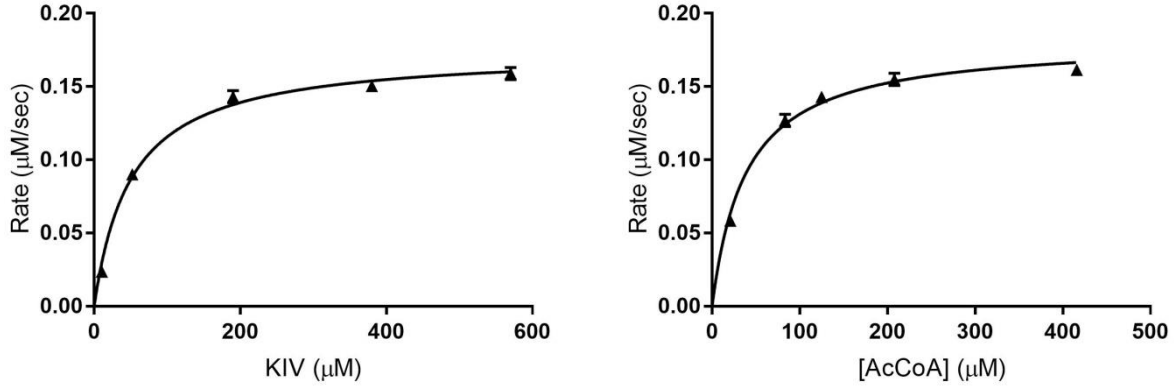


Figure 2.20: Michaelis-Menten kinetics of *Nme*IPMS E298A.When obtaining the apparent K_m for one substrate, the other substrate was held at a saturating concentration of 250 μM.

The K_m values for AcCoA for *Nme*IPMS Glu298Ala have not changed substantially compared to the wild type *Nme*IPMS, suggesting that this mutation does not affect the interaction of the protein with AcCoA (Figure 2.20, Table 2.5). The K_m for KIV has increased, from 36 ± 3 for the wild-type protein to 51 ± 5 for the Glu298Ala mutant. As Glu298 is in subdomain I, which contributes residues to the active site, altering the interactions that subdomain I makes with the barrel or within the active site may cause an increase in K_m for KIV. The k_{cat} has decreased compared to the wild type protein, causing a substantial decrease in k_{cat}/K_m. Although this mutation does impact catalysis, there is a less than two-fold reduction in activity, suggesting that the residue does not contribute significantly to catalysis, but may play a role in anchoring subdomain I to the catalytic barrel to facilitate catalysis. The mutation of Glu298 to Ala has a similar effect on catalysis to the Arg470Ala and Arg32Ala mutations.

2.3.2.3 Leucine inhibition of *Nme*IPMS Glu298Ala

The IC_{50} for leucine for wild type *Nme*IPMS and *Nme*IPMS Glu298Ala were determined (Figure 2.21, Table 2.5). The IC_{50} for leucine for wild type *Nme*IPMS was similar to previous characterisations of untagged protein, suggesting that the addition of the His_6 tag does not adversely impact leucine inhibition.[1, 64, 130] The IC_{50} was determined at saturating AcCoA and KIV substrate concentrations.

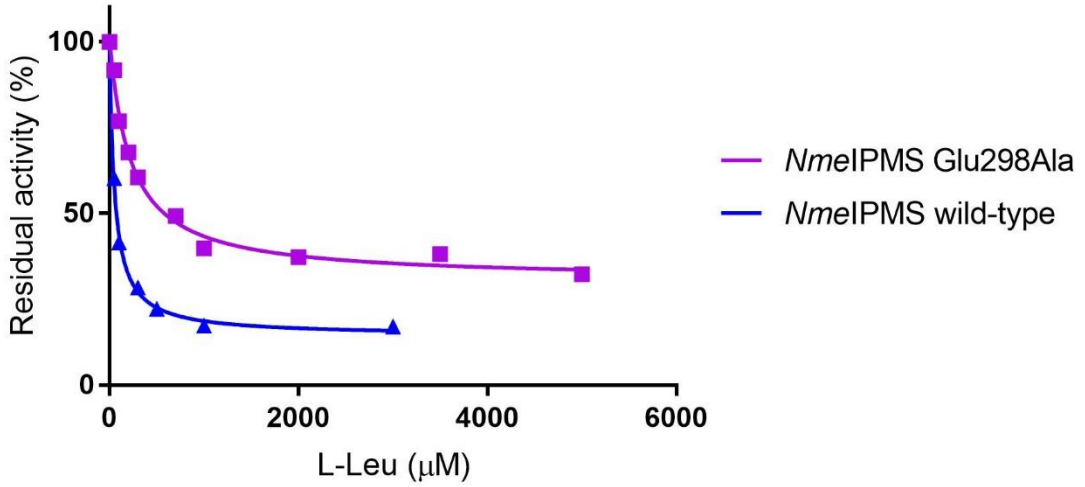


Figure 2.21: The IC_{50} for leucine for *Nme*IPMS wild-type, and *Nme*IPMS Glu298Ala.The IC_{50} for *Nme*IPMS wild-type (blue) and *Nme*IPMS Glu298Ala (purple) were determined at an AcCoA of 230 μM and a KIV concentration of 210 μM.

The IC_{50} for leucine for *Nme*IPMS Glu298Ala has increased from 50 μM to 220 μM ± 20 μM. This analysis was performed at saturating AcCoA and KIV substrate concentrations, as with the wild type protein IC_{50} determination. This suggests that although the mutation has a limited effect on catalytic activity, there is an almost four-fold increase in IC_{50}, suggesting that the protein is less inhibited by leucine than its wild type counterpart. Additionally, the residual activity for the wild type protein, at a high concentration of leucine, has been determined at between approximately 14% residual activity. To further investigate the inhibition profile of the *Nme*IPMS Glu298Ala mutant, the effect of leucine at several concentrations while the substrate is varied could be investigated. This was not performed due to time restraints.

The residual activity for the Glu298Ala mutant is higher than that of the wild type protein at 30%, suggesting that inhibition by leucine is less complete in this mutant compared to the wild type. De Carvalho et al.[63] determined the inhibition by leucine of a number of mutants in *Mtu*IPMS, and *Mtu*IPMS His379Ala showed increased residual activity in the presence of leucine compared to the wild type protein although there was no change in the K_i or K_i* values for this mutant. The authors suggested that the binding of L-leucine was not affected by this mutation, but the active site was less altered in the mutant when leucine is bound compared to the wild type protein, allowing for an increase in residual activity.

Product release is thought to be a key, and possibly rate determining, step in the IPMS reaction cycle, as this may involve a large-scale movement of the subdomains to facilitate release of products and subsequent binding of substrates.[56] The asymmetry noted in the *Mtu*IPMS structures is suggestive of a 'closed' and 'open' active site, and MD simulations on *Nme*IPMS have suggested that AcCoA can be more readily recruited to the open active site. Therefore, the increased residual activity may be because the active site may not remain closed efficiently as Glu298 is in subdomain I and forms an interaction at the top of the barrel that may help to anchor subdomain I to the top of the barrel. In the absence of this interaction when Glu is mutated to Ala, substrate binding and release may occur more readily in the presence of the inhibitor than in the wild type protein.

Although the Glu298Ala mutation did not abolish leucine sensitivity as the Arg470Ala and Arg32Ala mutants did, it did have a substantial impact on inhibition by leucine. This suggests that this residue may form part of the allosteric network that alters the active site in response to leucine.

Table 2.5: A summary of the kinetic and inhibition parameters of *Nme*IPMS wild type and several alanine mutants. * denotes parameters determined by Plowman-Holmes[131]. N/A stands for not applicable.

Protein	K_m (KIV) μM	K_m (AcCoA) μM	k_{cat} (s^{-1})	L-Leu IC_{50} (μM)	Residual activity (%)
Wild type *Nme*IPMS	36 ± 3	35 ± 3	7.2 ± 0.1	53 ± 5	17
*Nme*IPMS R470A*	55 ± 3	58 ± 6	2.8 ± 0.1	N/A	N/A
*Nme*IPMS R32A*	39 ± 4	44 ± 4	3.1 ± 0.1	N/A	N/A
*Nme*IPMS E298A	51 ± 5	39 ± 4	4.7 ± 0.1	220 ± 20	30

2.4 Summary

With the massive increase in sequence data available, techniques to identify patterns and provide meaning to that data are becoming increasingly available and sophisticated. Covariance analyses, such as statistical coupling analysis, present a way to explore protein evolution in a new light. Such techniques do require knowledge of the system they are analysing, and structural and functional information is required for further investigation of the system. Additionally, these types of analyses require a thoroughly curated multiple sequence alignment, and, particularly for SCA, an understanding and analysis of the phylogenetic basis of the sequence population. If little is known about the system, it can be difficult to produce an accurate sequence alignment, especially if the alignment contains regions with significant gaps.

Covariance analyses provide a way to explore sequence space that is not reliant on absolute conservation. Residues that may not be conserved but are important for protein function can be readily identified, and aspects of protein behaviour such as protein-protein interaction, functional dynamics, and allosteric regulation, can be explored more broadly if the pathways that underline these aspects can be identified and modulated. Covariance analyses also provide opportunities to explore enzyme evolution in the context of drug resistance, as identifying regions of the protein that are key for catalysis or allosteric regulation that are not binding sites may provide additional targets or ways to modulate functionality and suggest how organisms have evolved, or may evolve, to evade drugs.

The *Nme*IPMS-like IPMS were chosen as the target for this investigation, as this group of enzymes provide an interesting picture of a very dynamic protein in which the mechanism of allostery is not well understood. Also, it is difficult to study this protein by purely biophysical or structural means as it does not readily crystallise, and protein stability can also be problematic. Therefore, covariance analyses provide a different way to investigate the important residues controlling protein motion that facilitate catalysis and allosteric regulation. As IPMS is found almost invariably in bacterial species, a wide phylogenetic net can be cast to limit the impact of a narrow phylogenetic or environmental niche on the covariance analyses that may provide less information about the protein motion as a whole.

The SCA identified one sector that spanned from the N-terminal catalytic domain to the C-terminal regulatory domain of *Nme*IPMS. The sector contained numerous residues in the dynamic subdomains, as well as ones surrounding the leucine binding site, suggesting a potential role for this sector in the transmission of allostery through the subdomains that are so critical for AcCoA binding to the catalytic domain.

In light of the identification of this sector, several mutations were made in *Nme*IPMS to assess the character of this pathway. Some of these residues had also been identified through other means, such as MD simulations, that provide further evidence for the importance of these residues in protein function. One mutation, Glu298Ala, had a limited impact on catalytic activity, but attenuated allosteric inhibition substantially, while two other mutations, Arg470Ala and Arg32Ala, also had limited impact on catalysis, but completely abolished allosteric inhibition by leucine. Both Arg470Ala and Arg32Ala did still bind leucine, as evidenced by isothermal titration calorimetry. This evidence suggests that residues in this SCA sector do contribute to the allosteric network. It also suggests that SCA, as discussed previously, may provide a way to identify these networks that are not apparent by simple multiple sequence alignment, is not reliant on structural information,

and is less expensive computationally than MD simulations. SCA also does not rely on structural information.

This analysis demonstrates the power of covariance analyses such as SCA to identify networks that are not apparent from other types of analyses. It provides additional insight into the complex mechanism by which this group of enzymes perform their function and how this catalytic activity is regulated, and how allostery may evolve in a narrow phylogenetic group.

Chapter 3: Covariation analysis of IPMS, CMS, and HCS, from bacteria and archaea

3.1 Introduction

The statistical coupling analysis of the *Nme*IPMS-like IPMS group demonstrates that this type of analysis can provide promising leads and valuable information about this very dynamic system. However, Kumar et al.[69] suggest that there are evolutionarily distinct versions of IPMS, CMS, and HCS (Figure 3.1). What was termed IPMS1 is characterised by IPMS from *Methanocaldococcus jannaschii* (*Mja*IPMS) and *Nme*IPMS, while IPMS2 was typified by *Mtu*IPMS. IPMS1 are phylogenetically close to CMS1, including *Mja*CMS, while *Lin*CMS-like CMS fall into a discrete group termed CMS3. The HCS group also divides along phylogenetic lines. As different members of these groups can show allostery affecting different parts of the catalytic cycle, as discussed in Chapter 1, it suggests that the mechanism of allostery may be tied to phylogeny, or at least is not conserved through all the diverse groups. Therefore, to explore the underlying mechanisms of catalysis and regulation more widely, a broader phylogenetic approach must be taken.

One interesting feature of this broad group of proteins is the presence and absence of the regulatory domain. Additionally, the subdomains appear to be of critical importance in the recruitment of AcCoA and facilitation of catalysis in both the regulatory domain present (RDP) and the regulatory domain absent (RDA) proteins. This suggests that there must be coevolution of residues to facilitate sufficient flexibility and stability in the subdomains to allow for catalysis in the presence or absence of a regulatory domain.

There are similarities in structure between the two populations that suggest that they have a similar mechanism for binding AcCoA. The loop that connects the subdomains contains residues that are critically important for AcCoA recruitment, although Okada et al.[106] suggest that the groove in which the adenosine portion of AcCoA is proposed to bind in *Thermus thermophilus* HCS (*Tth*HCS) is a different shape to that in *Leptospira interrogans* CMS (*Lin*CMS), suggesting that there is potentially a difference in absolute substrate recognition between the two populations.

Therefore, coevolution analyses were used to probe these broad protein populations to investigate whether there is coevolution in the subdomains that may facilitate catalysis in the presence or absence of a regulatory domain.

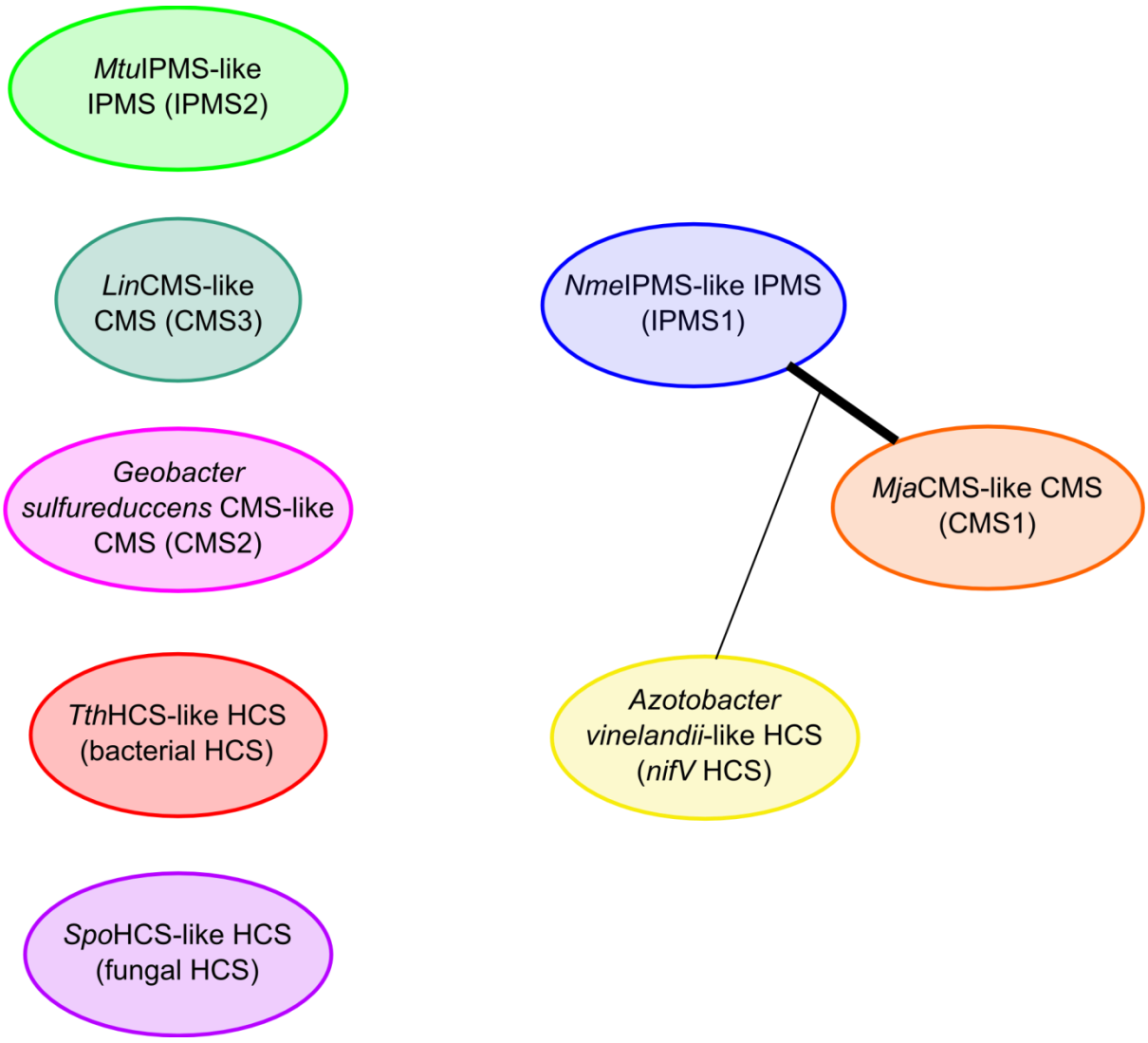


Figure 3.1: Some of the groups of IPMS and IPMS-like proteins defined by Kumar et al..[69]

3.2 Statistical coupling analysis of Claisen condensation-like enzymes

3.2.1.1 Multiple sequence alignment construction

A pool of sequences containing the LeuA Dimer motif was obtained from Pfam.[114] As with the *Nme*IPMS-like sequence pool described in Chapter 2, sequences were removed from the population if they were less than 300 amino acids, greater than 700 amino acids, or lacked key conserved residues or motifs. An initial multiple sequence alignment of 888 sequences of *Nme*IPMS-like IPMS, *Lin*CMS-like CMS, and *Mja*CMS-like CMS sequences was obtained using MAFFT, seeded by a structural alignment performed by PROMALS3D.[92, 134]

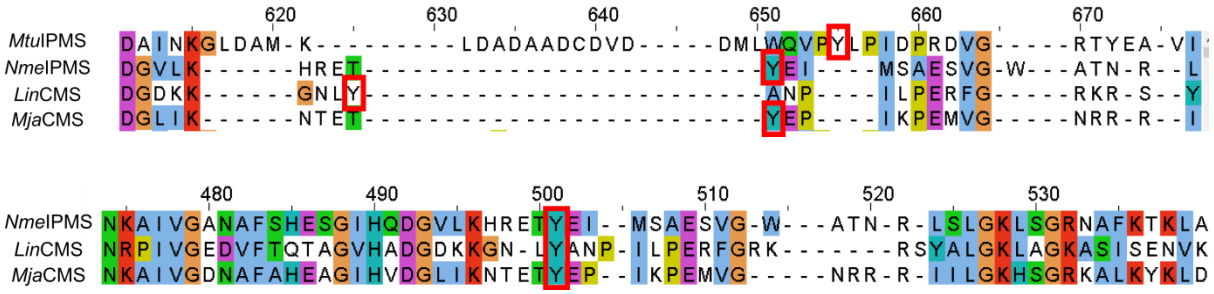


Figure 3.2: MSAs showing the region around the conserved tyrosine in subdomain II known to be important for catalysis. The tyrosine of interest (Tyr410 in *Mtu*IPMS, Tyr313 in *Nme*IPMS) is shown in the red boxes. When *Mtu*IPMS and related species are present in the alignment, this region is mis-aligned in this region, while when *Mtu*IPMS-like IPMS are not present, this region is not mis-aligned.

*Mtu*IPMS-like IPMS sequences were not used in this analysis due to problems with producing an accurate alignment (Figure 3.2). The *Mtu*IPMS-like IPMS sequences form a separate cluster in the protein similarity network constructed by Kumar et al. and in the CLANS analysis discussed in Chapter 3.2.1.1.1.[69] The length of the IPMS-like IPMS sequences is also considerably longer, for example, the average length of the *Mtu*IPMS-like IPMS sequences was approximately 590 amino acids while *Nme*IPMS-like IPMS sequences have an average length of approximately 515 amino acids. The structural alignment between the *Lin*CMS and *Mtu*IPMS structures was inaccurate as the extension in the middle of the *Mtu*IPMS sequence and the lack of structural information for subdomain I and II in the *Lin*CMS structures meant that the MSA was mis-aligned if *Mtu*IPMS-like IPMS sequences were present.

The structural alignment was performed using the PROMALS3D web-server, with the homology model of *Nme*IPMS, and one of the structures of *Lin*CMS (PDB: 3BLF) as input.[134] The RMSD of the *Nme*IPMS homology model and the partial structure of *Nme*IPMS (PDB: 3RMJ) is 1.96 Å,

suggesting that, based on the structural evidence available, it provides an adequate representation of the protein structure. As the structure of *Lin*CMS is not complete, this structural alignment only aligned the catalytic barrels of both proteins. This structural alignment was then used as a 'seed' for a large-scale multiple-sequence alignment in MAFFT. If the structural alignment was not used as a seed, the multiple sequence alignment was mis-aligned based on a comparison with a structural alignment produced in PyMol using the super alignment tool.

MAFFT is a multiple sequence alignment programme that utilises fast Fourier transform (FFT) to perform alignments.[92, 127] The structural alignment produced by PROMALS3D was used as a 'seed'. This means that the aligned 'letters' (i.e. the residues) are preserved, but gaps are not. The additional sequences were added as full sequences, and MAFFT FFT-NS-2 was used to produce a final multiple sequence alignment. Due to the nature of the gap penalties with this algorithm, and the lack of structural information at this position from the *Lin*CMS crystal structures, the conserved Tyr in subdomain I that is equivalent to *Nme*IPMS Tyr313 in the *Lin*CMS-like sequences was manually aligned. This final, curated, alignment of 888 sequences was then used as input for SCA.

Alternative methods of sequence alignment were also considered. PRANK, a phylogeny-aware sequence alignment method, the EXPRESSO variant of T-COFFEE, and DECIPHER, a new multiple sequence alignment method that uses secondary structure prediction and gap penalties relevant to the local environment of each amino acid, were also tested for this sequence pool.[135-137] However, MAFFT, when seeded with the PROMALS3D structural alignment of the known enzymes, produced a reliable alignment as alignments produced by other methods mis-aligned conserved regions known to be important for catalysis.

As the output of the SCA is absolutely dependent on the quality of the alignment, an external method of validating the alignment was required. The LoCo tool was also used to assess the multiple sequence alignment by analysing local co-variance in the alignment.[138] There was limited local co-variance at most positions, aside from ones with biological relevance, such as those conveying specificity of the ketoacid substrate to the active site, suggesting that there was not significant mis-alignment within the alignment.

Pfam was also mined for a sequence population of Claisen condensation enzymes that did not contain regulatory domains. This sequence population was then manually edited to remove aberrant sequences and similar enzymes that bind other substrates, based on criteria established by Casey et al.[68] Fungal HCS sequences were not used in this alignment as they formed a discrete

phylogenetic group. A structural alignment between a structure of *Lbi*IPMS (PDB: 4OV4), an IPMS lacking a regulatory domain, and *Tth*HCS (PDB: 2ZTJ), a bacterial HCS, was performed using PROMALS3D and this was used to seed a multiple sequence alignment, performed in MAFFT. This alignment of 488 sequences was then manually assessed and used as the basis for the SCA of the RDA alignment.

3.2.1.1.1 CLANS of the sequence populations

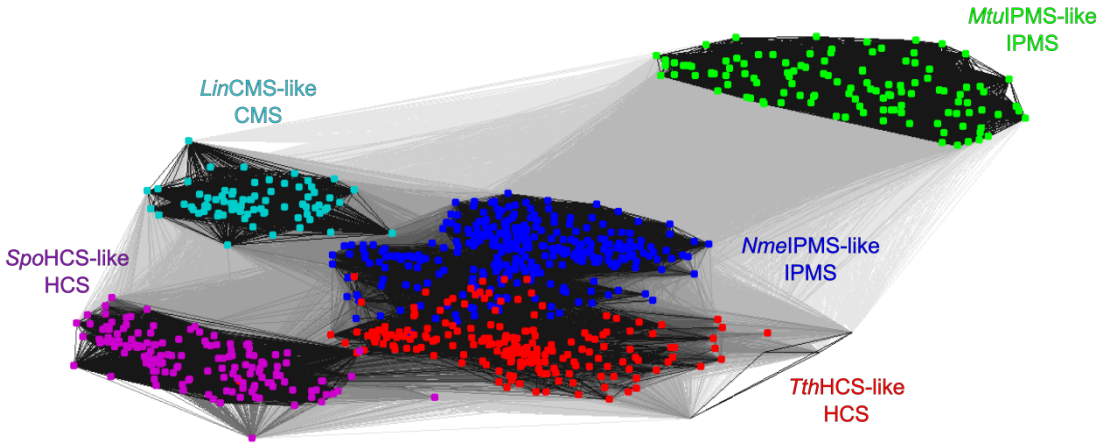


Figure 3.3: CLANS analysis of the sequence populations of interest including the *Mtu*IPMS-like IPMS sequences (green), *Nme*IPMS-like and *Mja*CMS-like IPMS and CMS (blue), *Tth*HCS-like HCS (red), *Spo*HCS-like HCS (purple), and *Lin*CMS-like CMS (teal).

All sequences obtained for this analysis were investigated using CLANS (Figure 3.3). This included *Mtu*IPMS-like IPMS sequences and fungal HCS sequences as CLANS does not require a MSA as input. This showed a similar pattern to that observed by Kumar et al.[69] with *Mtu*IPMS-like IPMS forming a discrete group away from the other sequences.

CLANS was also performed using only the RDP sequence population, and this demonstrated that there are three distinct groups in this population, but they form tight connections with one another (Figure 3.4).

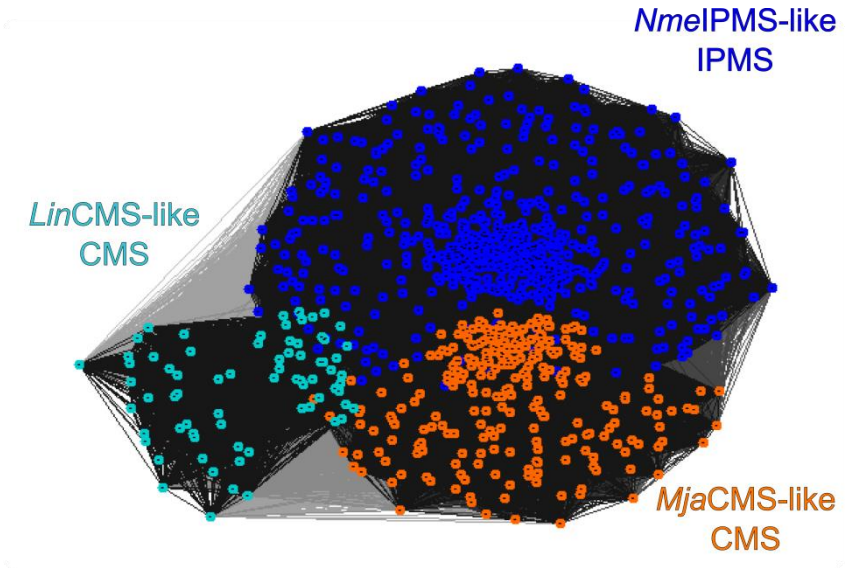


Figure 3.4: CLANS analysis of the RDP sequence population.The definition between the *Nme*IPMS-like IPMS (blue) and the *Mja*CMS-like CMS (orange) population is shown although it is absent in Figure 3.2. The *Lin*CMS-like CMS population forms a separate cluster.

3.2.2 Regulatory-domain present sequence SCA (RDP-SCA)

3.2.2.1 Sequence similarity in the RDP sequence alignment

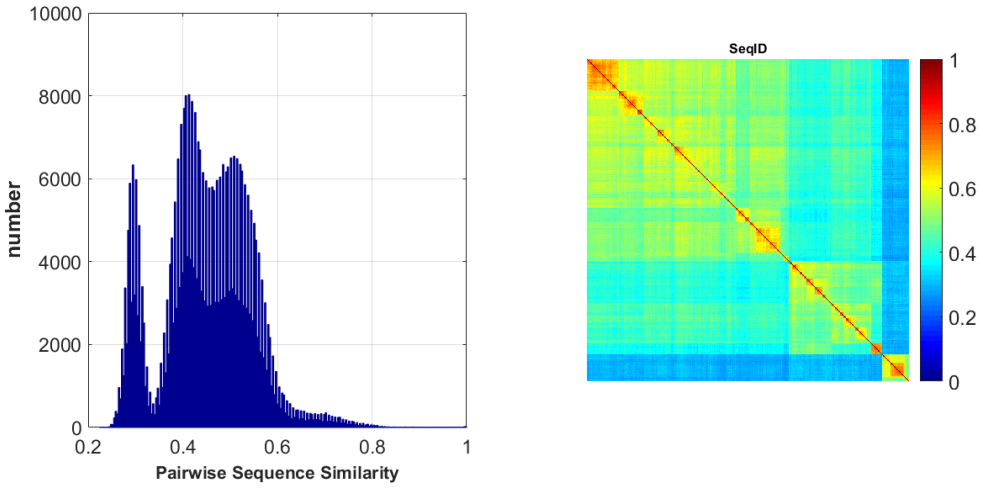


Figure 3.5: Sequence similarity in the RDP alignment from a matrix of similarity.

Sequence similarity in the RDP sequence alignment was assessed (Figure 3.5). The heat map (Figure 3.5, right) shows a matrix of similarity, the fraction of amino acids that are the same between two sequences. This shows that the sequence population is not homogeneous, as would be expected with a population of proteins that bind different ligands.

3.2.2.2 Conservation in the RDP sequence alignment

The positional conservation was also assessed by determining the relative entropy of each position (Figure 3.6). This shows that there is considerable conservation at similar positions to that of the *Nme*IPMS-like IPMS sequence population, for example the conserved histidine residues His204 and His206 important for metal ion coordination. There is less conservation overall in the RDP sequence alignment compared to the *Nme*IPMS sequence alignment, which correlates with the increased phylogenetic distance between the sequences of interest.

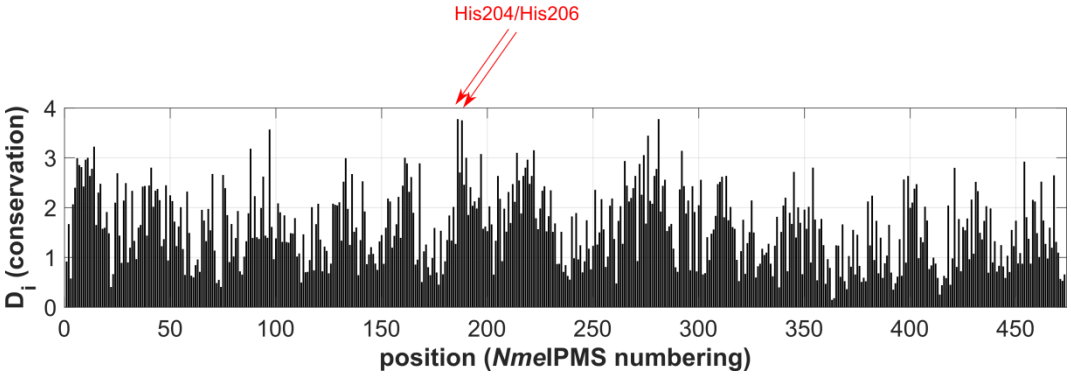


Figure 3.6: Positional correlation in the RDP alignment.Highlighted are absolutely conserved residues known to be important for catalysis. The residue numbers are from *Nme*IPMS.

3.2.2.3 SCA-PCA of the RDP sequence alignment

The positional correlation matrix of the RDP alignment was calculated using SCAv5.m as described in Chapter 2. The matrix (Figure 3.7) demonstrates the small amount of absolute conservation in the matrix (dark blue) but shows that there is considerable positional correlation spread through the matrix.

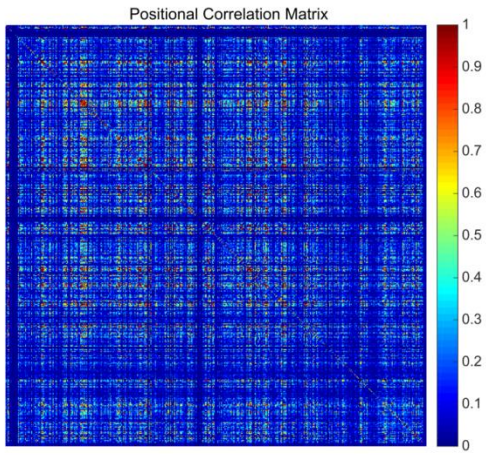


Figure 3.7: Positional correlation matrix from the SCA calculation of the RDP alignment.The colour gradient shows the degree of statistical coupling from blue (low) to red (high).

The eigenspectrum was also plotted (Figure 3.8). As described in Chapter 2, the red line denotes the randomised alignments that allow determination of the significant eigenmodes. The top three eigenmodes are statistically significant, and the top eigenmode has a substantially larger eigenvalue than the others, as observed in the *Nme*IPMS SCA. The top eigenmodes were also assessed by scatter plot. The scatter plots suggest that there are potentially at least two 'independent' sectors within the data (Figure 3.10). As it was known that multiple different phylogenetic groups were used in the construction of the alignment, independent component analysis (ICA), as opposed to the PCA used in the *Nme*IPMS SCA, was utilised to investigate the identity of these sectors.

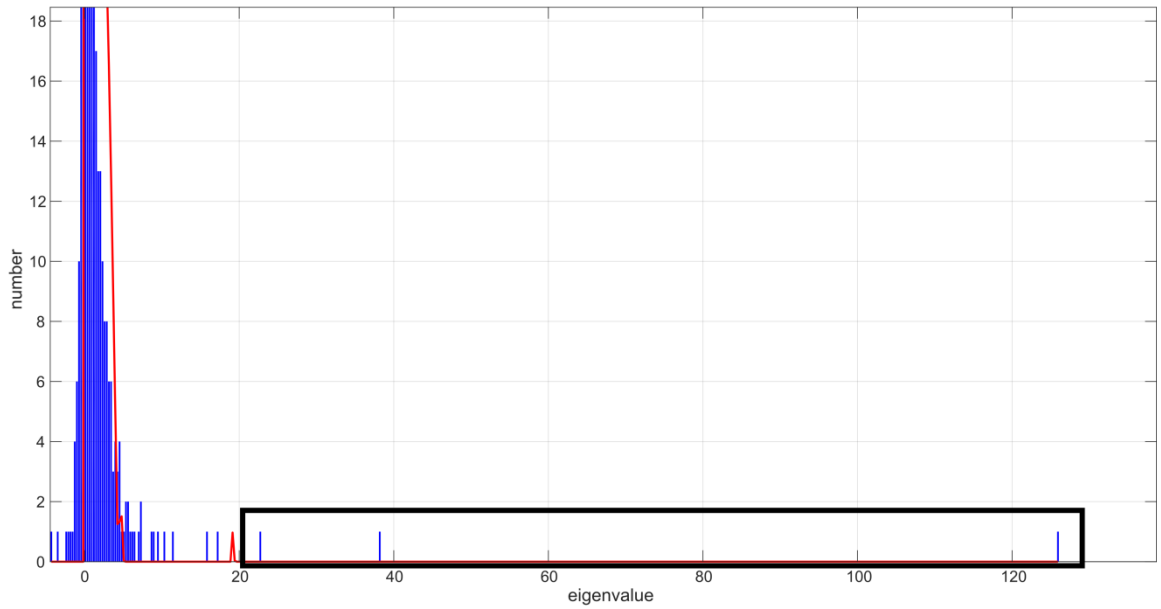


Figure 3.8: The eigenspectrum of the matrix produced by SCA after PCA for the RDP alignment.The black box indicates those eigenmodes of interest. The red line indicates the distribution of randomised matrices.

3.2.2.4 Independent component analysis (ICA) of the RDP-SCA

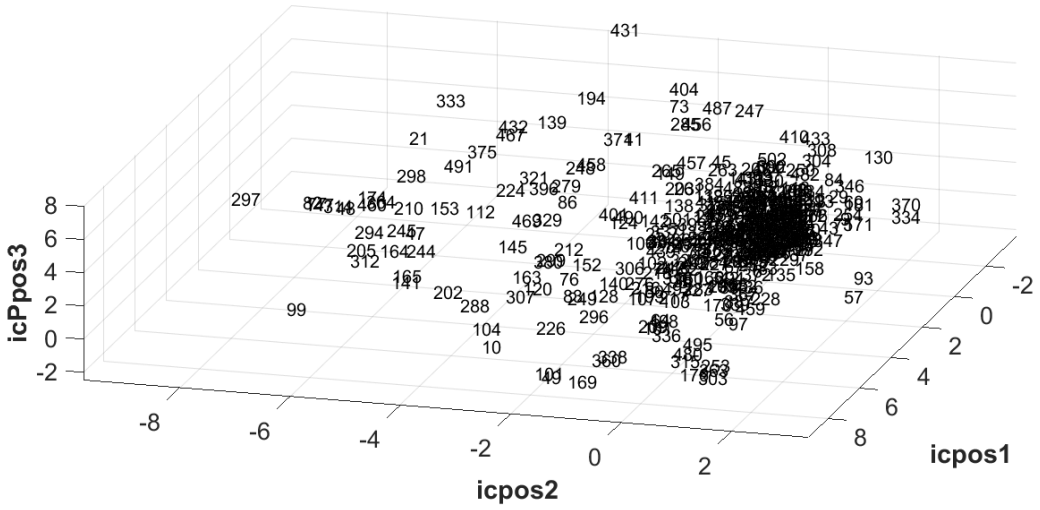


Figure 3.9: The top three independent components of the RDP SCA.

Independent component analysis (ICA) is a way of repositioning the data to explore the presence or absence of independent sectors within the SCA matrix. Principal component analysis determines the independent variable that best describes all of the data, while independent component analysis is used to extract the underlying components involved in producing the data. An example commonly used to describe independent component analysis is the "cocktail party problem" where several people are talking at once, so the resulting principal component is a mixture of all of the voices. If independent component analysis is used, the noise can be separated into individual voices again. In this analysis, the top 3 eigenmodes were selected for analysis, and the eigenmodes were transformed into independent components (Figure 3.9, Figure 3.10). Singular value decomposition of the sequence correlation matrix from the initial SCA was also performed to allow for identification of independent sectors that also displayed sequence correlation, i.e. sectors defined by a phylogenetic basis.

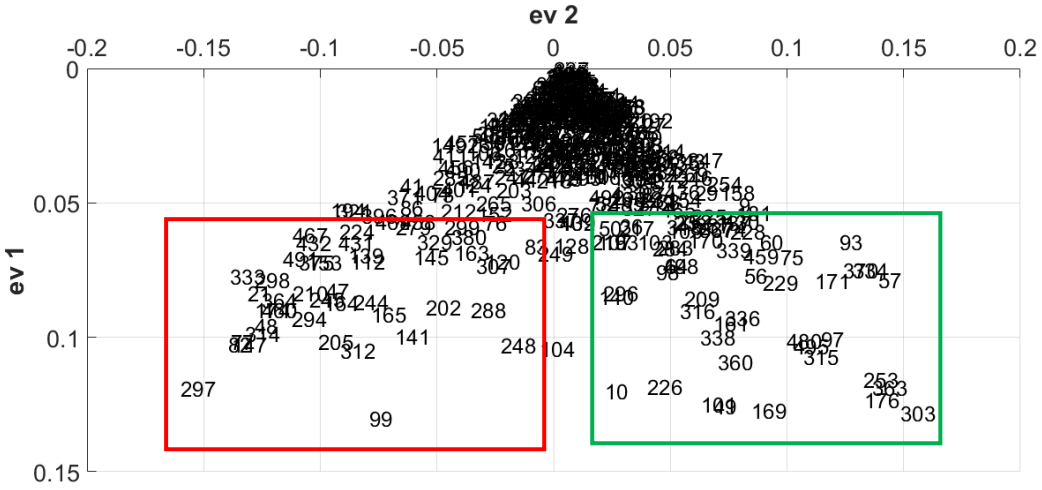


Figure 3.10: Scatter plot of the first and second eigenmodes demonstrating the potential two sectors within the data in the red and green boxes

When the top three independent components are plotted and compared to the pattern of sequence divergence, it is obvious that the residues with a high score for the second independent component (IC2) is due to the phylogenetic relationship in the sequences (Figure 3.11). Based on the multiple sequence alignment, this cluster of sequences is likely from the *Lin*CMS-like CMS group of sequences, and this reflects the CLANS analysis that clustered these sequences apart from the *Nme*IPMS-like IPMS group and the *Mja*CMS-like CMS group.

Upon further analysis of the residues in this IC, it appears that IC2 includes residues that are involved in AcCoA interaction. IC2 was also mapped onto the *Nme*IPMS homology model (Figure 3.12). This IC includes residue Arg77 (*Nme*IPMS numbering), which corresponds to Phe83 from *Lin*CMS. In the AcCoA-containing structure of *Lin*CMS (PDB: 3BLI), Phe83 a forms hydrophobic interaction with AcCoA in the *Lin*CMS structure, and Phe is conserved in the *Lin*CMS-like CMS sequences at this position, whereas Arg is conserved in the *Mja*CMS/*Nme*IPMS population in the alignment. Additionally, this IC contains residues such as Lys112, Lys364, and Arg371 that have been predicted to be involved in AcCoA interaction with *Nme*IPMS by docking studies performed by Dr. Wanting Jiao (personal communication, May 2014). This suggests that *Lin*CMS may have a different way of binding AcCoA to the other RDP proteins, and other residues that may be involved in AcCoA interaction show statistical coupling in these populations to account for this change.

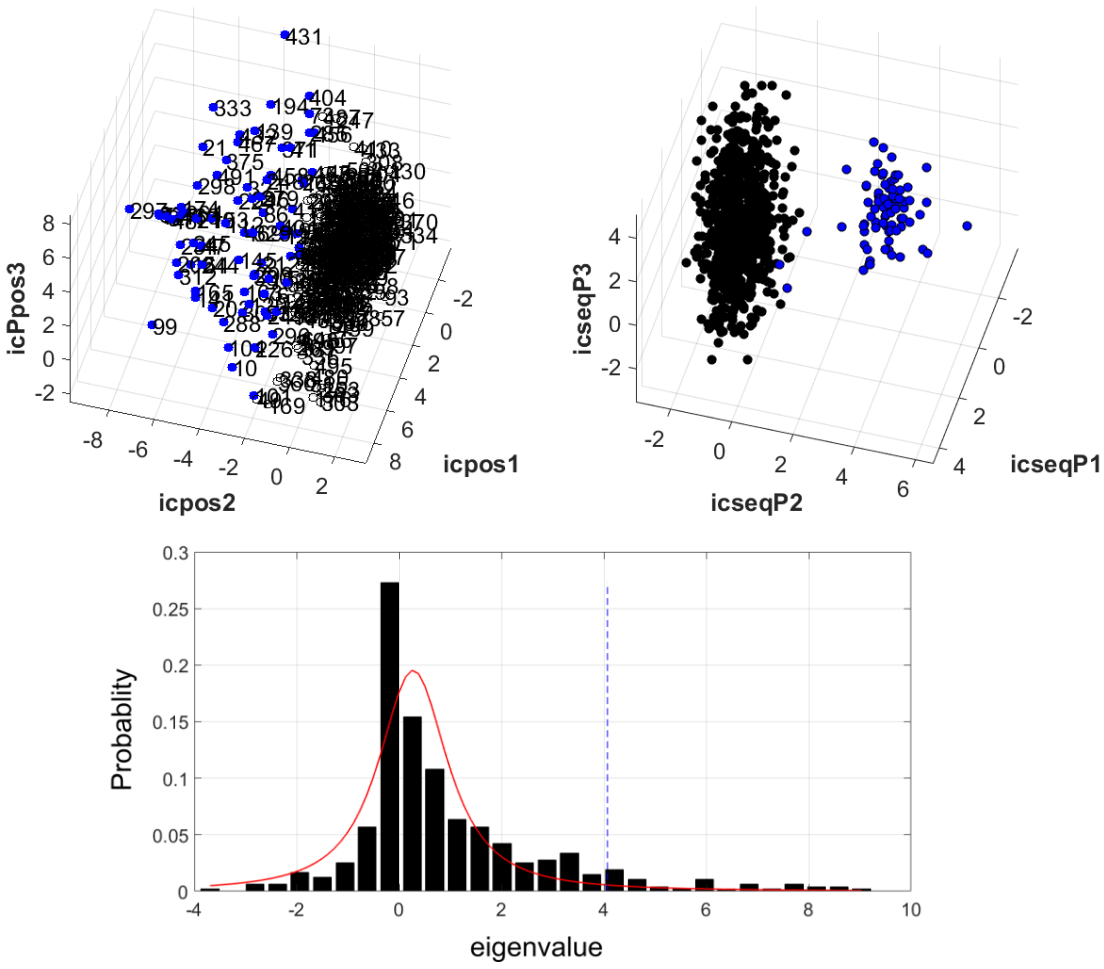


Figure 3.11: Scatter plots of the top three independent components (top, left), and the sequence space mapped to the independent component matrix (top, right). The blue spheres denote IC2 (top, left) and the sequence space that corresponds to IC2 (top, right). The histogram (bottom) indicates the scores for residues in IC2 plotted with a lognormal distribution (red line). The blue dashed line denotes the cut-off in the tail.

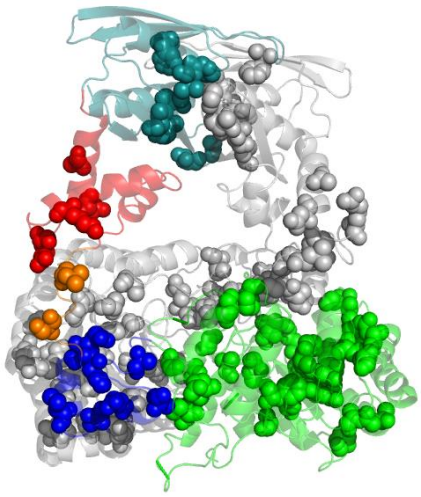


Figure 3.12: The residues with significantly high scores from IC2 mapped onto the homology model of *Nme*IPMS. Chain A is shown in grey, Chain B is coloured. The catalytic domain is shown in green, subdomain I in blue, the linker in orange, subdomain II in red, and the regulatory domain in teal.

Residues with a high score for the first IC (IC1) do not appear to have the same phylogenetic basis as IC2, as residues in IC2 showed statistical coupling due to the change in AcCoA interaction in the *Lin*CMS-like CMS group, and this was observed by comparing the sequence space with the independent components (Figure 3.11). The absence of a relationship with sequence suggests the residues in this independent component may provide more broad information about coevolved residues in the RDP population than IC2. IC1 contains residue Phe101, the only residue of those making contact with the ketoacid substrate that is different between all three broad phylogenetic groups involved in this analysis[69], further suggesting that this IC may give a broader picture of coevolved residues in these proteins than IC2.

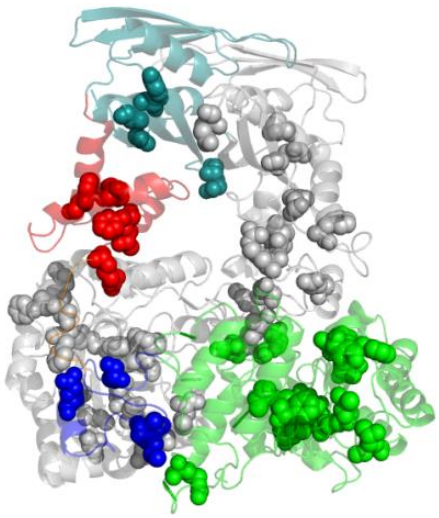


Figure 3.13: The residues with significantly high scores from IC1 mapped onto the homology model of *Nme*IPMS.Chain A is shown in grey, Chain B is coloured. The catalytic domain is shown in green, subdomain I in blue, the linker in orange, subdomain II in red, and the regulatory domain in teal.

When the residues in IC1 are highlighted on the homology model of *Nme*IPMS (Figure 3.13), they appear to show a network of interconnected residues that span the catalytic domain and subdomains, with a few residues in the regulatory domain. This suggests that this IC may represent a network of residues that are involved in controlling the motion of subdomain II to facilitate catalysis. IC1 also contains residue 336, which is an arginine in *Nme*IPMS. An alanine mutant of Arg336 in *Nme*IPMS has been made, and this change had a substantial impact on AcCoA interaction, with the K_m for AcCoA increasing to 800 μM.[64]

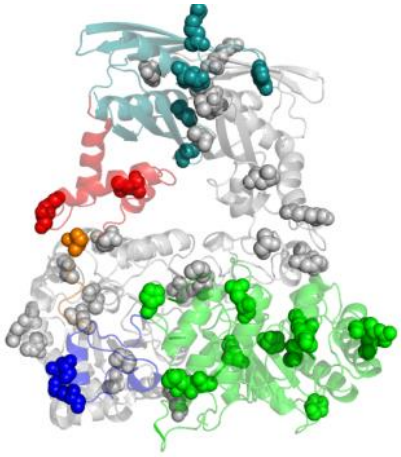


Figure 3.14: The residues with significantly high scores from IC3 mapped onto the homology model of *Nme*IPMS.Chain A is shown in grey, Chain B is coloured. The catalytic domain is shown in green, subdomain I in blue, the linker in orange, subdomain II in red, and the regulatory domain in teal

The third independent component of interest (IC3) has a number of residues with significant scores. These appear to be primarily located in regions of flexibility such as loops at the end of helices in the catalytic barrel, suggesting a potential role for the residues in this IC in maintaining overall structure and flexibility of the proteins (Figure 3.14).

3.2.3 Regulatory domain absent SCA

As with the RDP sequence population, the regulatory domain absent (RDA) sequence population was obtained from Pfam and was edited as described above. The fungal HCS sequences were removed as they skewed the sequence population significantly, so the sequence population, as with the RDP population, was made up of bacterial sequences. The fungal HCS sequences formed a distant group in the initial CLANS analysis (Figure 3.3), and when an alignment was constructed using these sequences, an initial attempt at using independent component analysis was performed. However, the differences between the bacterial and fungal RDA sequences were unable to be resolved. To investigate the similarities and differences between co-evolved residues in the two different populations, a further analysis using only fungal HCS sequences could also be performed although this was not done due to time constraints.

3.2.3.1 CLANS analysis of the RDA sequence population

CLANS analysis was also performed on the RDA sequence population to assess the sequence space (Figure 3.15). This suggested that, although there were several clusters, they were interlinked and there were no outlying clusters as seen in the RDP sequence population. The multiple sequence alignment was performed as described above.

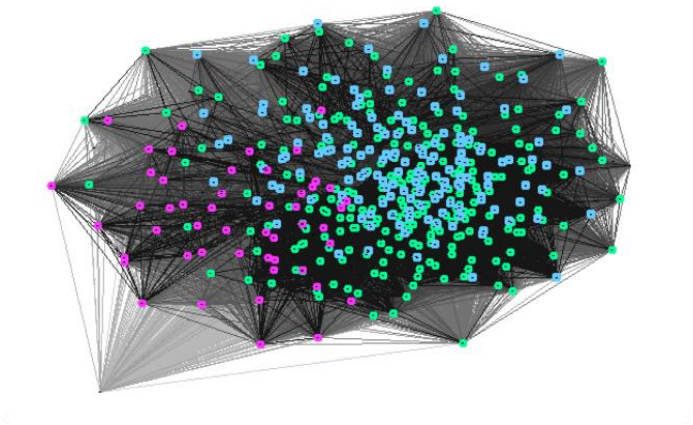


Figure 3.15: CLANS output for the RDA sequence populations.The different clusters identified by the program are shown in different colours. The clusters denote different clusters of proteins lacking regulatory domains. These clusters are not well defined by current definitions.

3.2.3.2 SCA analysis of the RDA alignment

This, along with analyses of the sequence similarity and positional correlation, was performed as described above for the RDP sequence alignment. The sequence similarity showed a similar pattern to that of the RDP alignment (Figure 3.16), although as with the CLANS pattern, there is closer sequence similarity amongst this RDA population than there is amongst the RDP population due to the outlying *Lin*CMS-like sequence group in the RDP population. As with the RDP sequence group, there is positional correlation at residue positions of significance, such as conserved histidines involved in metal ion coordination, and the conserved tyrosine in subdomain I (Figure 3.16).

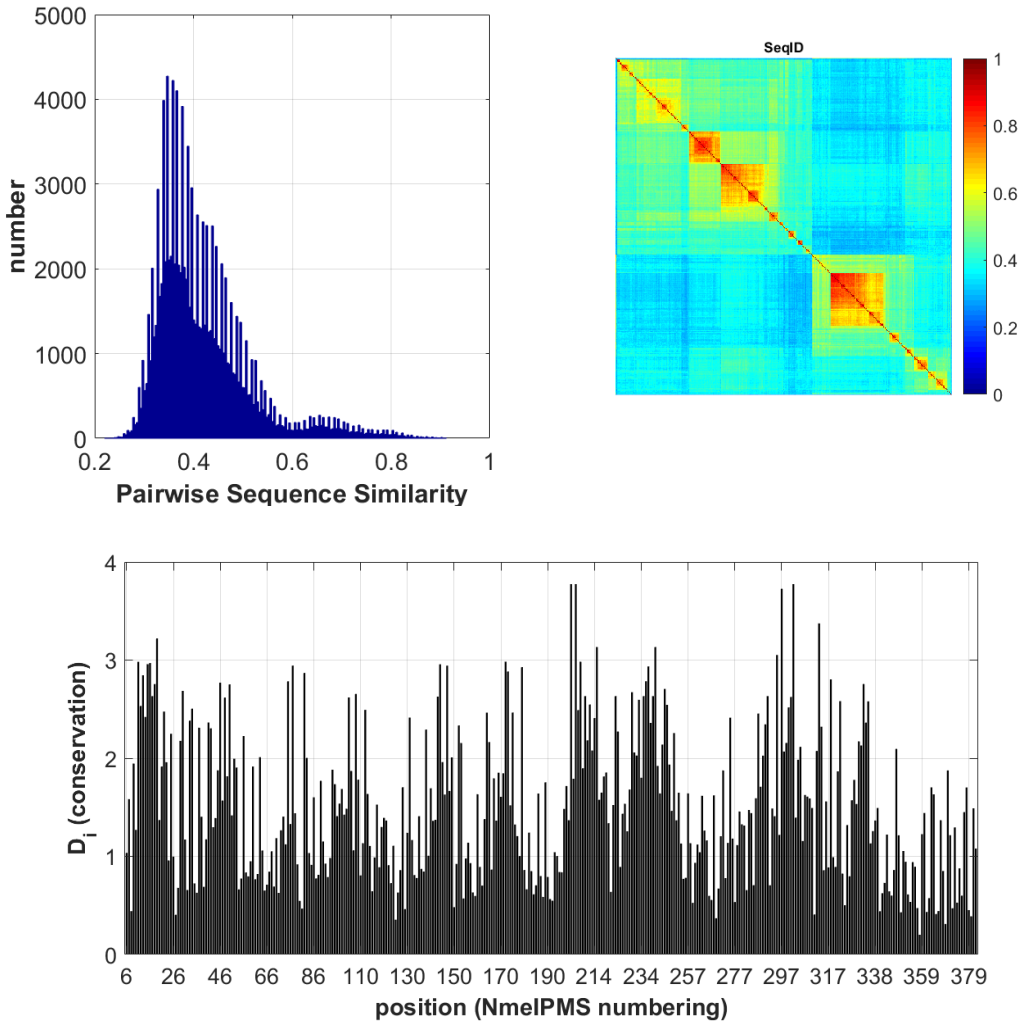


Figure 3.16: The sequence similarity in the RDA MSA (top) and the positional correlation of the residues in the alignment, *Nme*IPMS numbering.

The matrix obtained from the SCA calculations demonstrates a similar pattern to that of the RDP sequence alignment. Principal coupling analysis was performed on the matrix obtained from the SCA calculations, producing the eigenspectrum (Figure 3.17, top). There does not appear to be a substantial phylogenetic component to the coupling values observed in the matrix (Figure 3.17) so the top eigenmode was further investigated, as opposed to the RDP sequence alignment where there was an obvious phylogenetic component to the top eigenmodes.

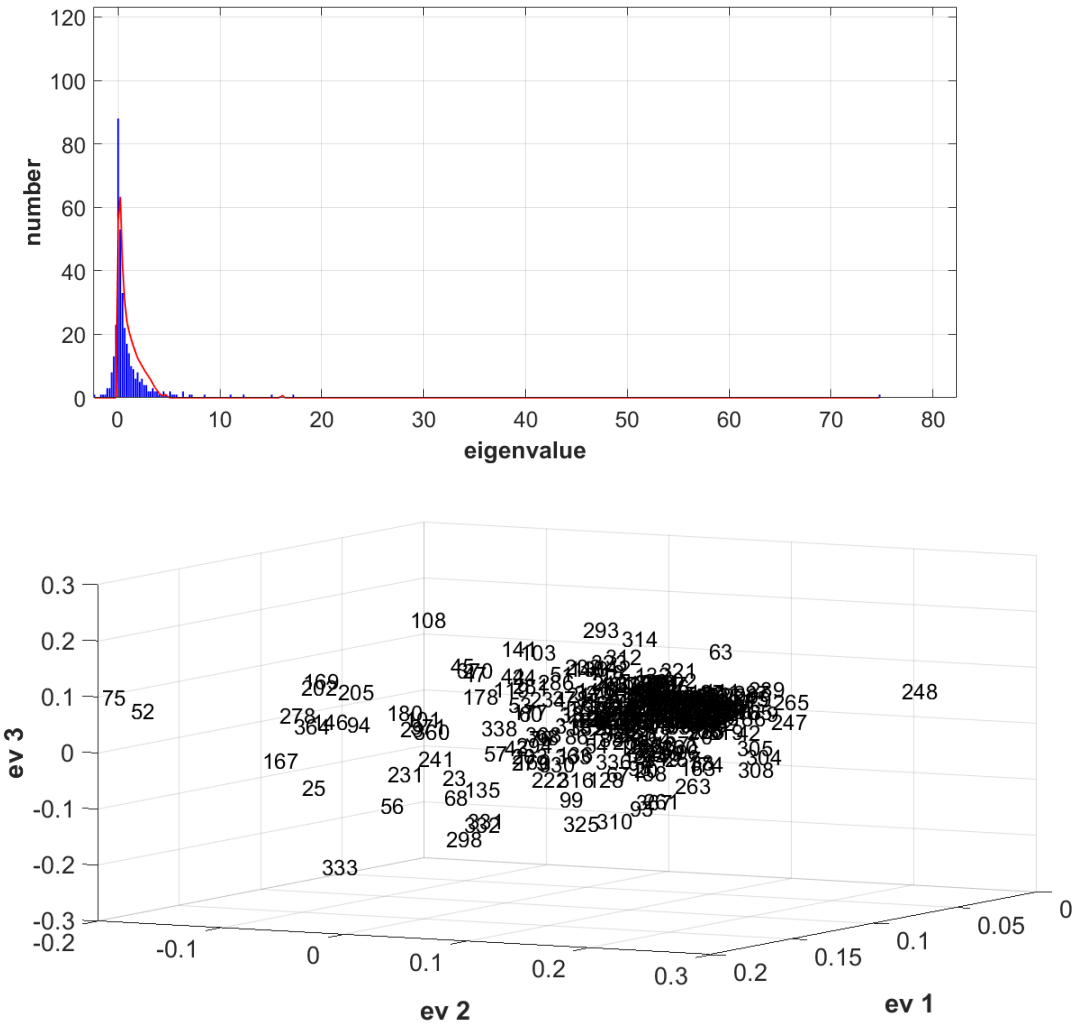


Figure 3.17: The eigenspectra (top) and the top three eigenmodes (bottom) of the RDA SCA.

The top eigenmode was fitted with a lognormal distribution, and the independent sector in the RDA matrix was defined by the construction of a cumulative density function and cutting off this so only residues in the tail were selected as sector residues. When these residues were plotted onto the homology model of *Nme*IPMS, they form a network spanning the subdomains and the catalytic domain (Figure 3.18). This network includes residues known to be involved with substrate

selectivity in the active site, such as Phe101, His99, and Leu75, (*Nme*IPMS numbering) suggesting that there may be a phylogenetic basis to the coupling observed or that when there is a change in these residues, there is a change in other residues to compensate for the altered shape or charge in the active site.

This sector also includes residues thought to be important for AcCoA interaction. The positive charge in the linker region between subdomain I and subdomain II is thought to be critical for the recruitment of AcCoA, and residues in this linker region show statistical coupling, suggesting that maintaining the charge in this position is important for facilitation of catalysis in the absence of the regulatory domain. Residue Glu298 was also identified in the *Nme*IPMS-like IPMS SCA (Chapter 2) but not in the RDP SCA analysis, suggesting that this residue may have different roles in the *Nme*IPMS-like IPMS population in allosteric regulation but yet not in facilitation of catalysis in the broader group, demonstrating the interconnectedness of residues involved in both catalysis and allosteric regulation in these proteins.

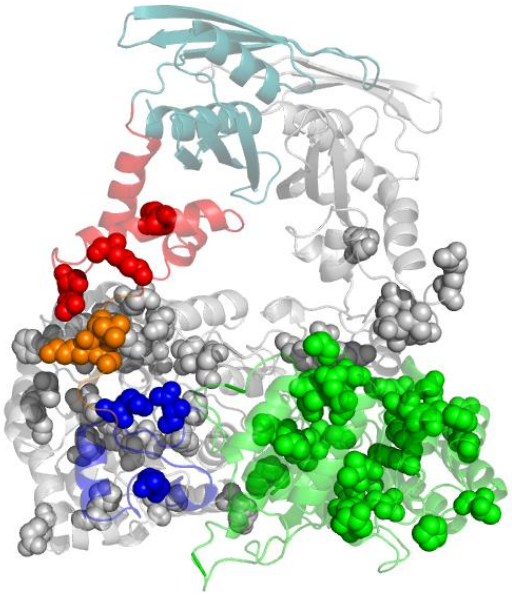


Figure 3.18: The network of residues identified as the sector by principal component analysis of the RDA SCA matrix.

There are several residues that also appear in the SCA of the RDP group. Phe360 (*Nme*IPMS numbering), located in subdomain II, shows statistical coupling in both the RDP and RDA SCA, implying that this residue has a critical role in catalysis in both populations regardless of the large structural change between the two populations. Phe360 in the RDP population is predominantly phenylalanine while in the RDA, this residue position is predominantly leucine, suggesting a

change from a large to small hydrophobic residue. In the *Nme*IPMS homology model, this residue is in the middle of the three-helix bundle that comprises of subdomain II (Figure 3.19, left).

A change in this position may substantially change the structure of the helices that make up subdomain II, that could stabilise or destabilise the subdomains, altering the recruitment of AcCoA and subsequently, catalysis. Asn169 (Figure 3.19, right), which also shows statistical coupling in both analyses, is found inside the active site, suggesting that a change in this position could be coupled to changes in the active site in response to substrate selectivity changes in the active site.

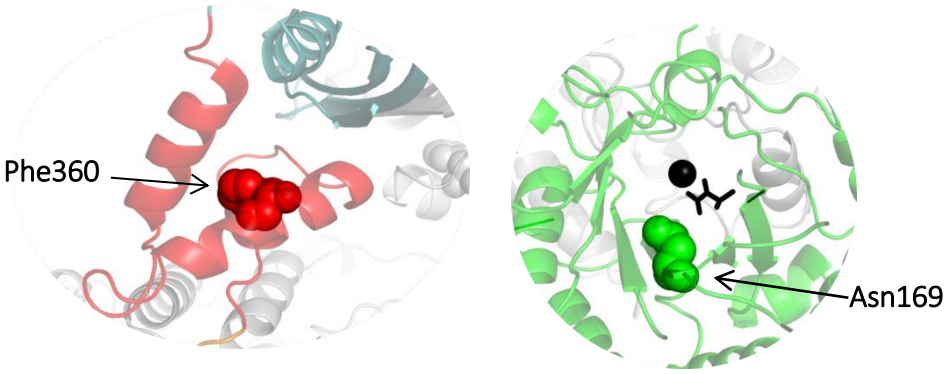


Figure 3.19: The location of Phe360 in subdomain II, and Asn169 in the active site of *Nme*IPMS. Phe360 is shown as red spheres, and Asn169 is shown as green spheres. The metal ion (sphere) and KIV (stick) are shown in black.

3.3 Summary

The comparison of the two statistical coupling analyses presents an interesting picture of these two groups of structurally similar proteins. There is a number of differences between the two populations. The RDP alignment could be best explored using independent component analysis as there was a phylogenetic bias to the alignment. This bias was not apparent in the RDA alignment, so principal coupling analysis could be used to investigate the presence of sectors. Both SCA produced sectors that spanned the catalytic domain and subdomains, suggesting coupled residues that control the motion of the subdomains are critical for catalysis. Very few residues in the regulatory domain in the RDP SCA were coupled, suggesting that there is not a conserved pathway of allosteric regulation amongst those proteins containing a regulatory domain. This correlates with differences in mechanism of allostery observed between the different populations.

There are a number of residues that show statistical coupling in both groups. Some of these residues are contained within the subdomains, suggesting that some residues are important for

maintaining the structure or dynamics of the subdomains in the presence or absence of the regulatory domain. As demonstrated by Phe360, there appears to be a difference in the size of residue in that position in the RDA compared to the RDP, suggesting that although this residue shows statistical coupling in both populations, there is also significant differences in the two alignments.

Some of these residues are involved in substrate selectivity, suggesting that in both proteins, there is possibly a functional basis to the coupling, although it is reasonable to presume that changes in the active site to accommodate other substrates are accompanied by changes in other regions of the protein regardless of specific phylogenetic changes. However, one way to explore this would be to use a method that is not influenced by phylogeny. As discussed above, the influence of the weighting of the SCA matrix by conservation can cause a bias in the subsequent analysis. If a substantial subset of the sequence population is from a different phylogenetic group and has significant changes in residues that are well conserved such as the *Lin*CMS-like CMS group, principal component analysis cannot be used to analyse that matrix. Although the RDA sequence population has substantial phylogenetic differences, ICA failed to identify these, suggesting that principal component analysis is sufficient to analyse this population.

This phylogenetic bias in SCA was also noted by Colwell et al.[139] who explored the statistical coupling analysis performed by Halabi et al.[129] on a group of serine proteases, and compared it to a covariance analysis performed by Skerker et al.[140] who used mutual information (MI) to investigate the interaction between histidine kinases and their response regulators. Colwell et al.[139] demonstrated that, although the algorithms used in the analyses were similar, they were not interchangeable in that if the MI algorithm was used on the serine protease sequence alignment, the results were considerably different and vice versa. They identified that the major difference between the two algorithms was the conservation weighting function used in SCA, and if this was used with the MI algorithm, then the results were comparable to the results from the SCA algorithm, and conversely, the SCA algorithm in the absence of the conservation weighting function produced similar results to the MI algorithm, which does not have a similar weighting function. This demonstrates clearly that the clusters identified by SCA are dependent on the conservation weighting function, and thus phylogenetic bias. This is particularly evident when there are changes in conserved residues and can skew the results significantly. Teşileanu et al.[141] also argued the same point, demonstrating that the functional sectors identified in prior SCAs may actually be identifiable through conservation, if single-site statistics are used to identify them.

As the residues identified by SCA in these two analyses present a potentially interesting picture of how these proteins may be able to bear the burden of the regulatory domain and still allow catalysis using the dynamic subdomains. However, the significant phylogenetic bias to the results by the conservation weight matrix used in the SCA becomes problematic as there are significant phylogenetic differences particularly in the RDP sequence population. Therefore, mutual information was also used to explore the same sequence alignments to compare and contrast the results as done by Colwell et al.[139] to explore the potential network of residues controlling the dynamics of the subdomains and catalysis.

3.4 Identification of covariance in IPMS and IPMS-like enzymes using mutual information

Mutual information, in this context, was adapted from information theory, and is a measure in the uncertainty about a position, in this case a residue in an MSA, given information about another position (residue).[142] It can vary from 0 to 1, with 1 indicating complete knowledge. MI is calculated based on the Shannon entropy (H), a measure of how random the residue population is in a particular column of a MSA.[143] Formerly, MI was a relatively imprecise way of detecting covariation in an MSA, as factors such as the number of sequences, the sequence diversity, and phylogenetic biases all hindered the usefulness of the algorithm.

Dunn et al.[143] developed a new metric, MIp, that utilises MI, but can accurately and quickly identify coevolved residues without the limiting factors of phylogeny. This technique uses a correction, termed the average product correction (APC), which determines the background mutual information of a particular alignment, which includes random noise and phylogenetic effects, and is used to correct that alignment so that only the residues that show mutual information related to structure and function, as opposed to background or phylogenetic aspects are identified. The difference between total MI, and MI when corrected, is presumed to be MIsf, or mutual information due to structural or functional restrictions on the protein, and thus are of interest. Dickson[144] then developed an efficient way of calculating mutual information using C and Perl-based programmes to calculate MI, MIp, and a variety of other statistics. This was implemented in Linux as the MIp Toolset. The MIp Toolset was utilised in the following analysis to obtain information about the mutual information of various groups of residues in alignments of subsets of the Claisen condensation-like enzymes.

3.4.1 Sequence populations and alignments

Buslje et al.[124] assessed the size of sequence alignments needed for effective predictability from the MI algorithm, with and without the average productive correction discussed above. They determined that an alignment size of over 400 sequences, where the sequences contained within are less than 62% identical, was required to provide sufficient numbers and variability of sequence to sufficiently reduce the impact of random noise and phylogenetic relationships so the output provides sufficient predictive performance. The RDP and RDA sequence alignments used in the

MIp analysis therefore remained the same as in the SCA analysis as both alignments fall within these parameters.

3.4.2 Mutual information analysis

The analysis was performed as described above, utilising the MIp Toolset that contains multiple programmes including the MIp.pl programme used for determining mutual information within a MSA. The output of this programme consisted of a large table of information, and a .dot file for easier viewing. One other programme within the MIp Toolset is dist_pdb that allows distances between Cα carbons to be determined from a PDB file. Chain A and Chain B from the *Nme*IPMS homology model, discussed previously, were independently used as an input for this programme.

Gloor et al.[145] noted that mutual information identified two types of co-evolving residues, residues that coevolved with only one or a small subset of other residues, termed single pairs, and those that form larger networks of residues and are typically at an active site or regions essential for protein function. They suggest that single pairs are typically important for local structure and the amino acid sidechains involved in such pairs are in contact, while the large group of residues are more likely to have a broader functional impact.

BioPython was used to produce contact maps for the *Nme*IPMS homology model, both within a chain and between the two chains. This allowed assessment of the single pair residue groups identified in the MIp analysis (Figure 3.22) to determine whether these single pairs showed the same pattern identified by Gloor et al..[145] Onto the *Nme*IPMS contact map, the residues with significant Z scores from either the RDP MIp analysis (Figure 3.20, top), or the RDA MIp analysis (Figure 3.20, bottom) were mapped. The single pairs of residues that share mutual information in both the RDP and RDA MIp analyses appear to be related to local structure, and also are typically close in sequence.

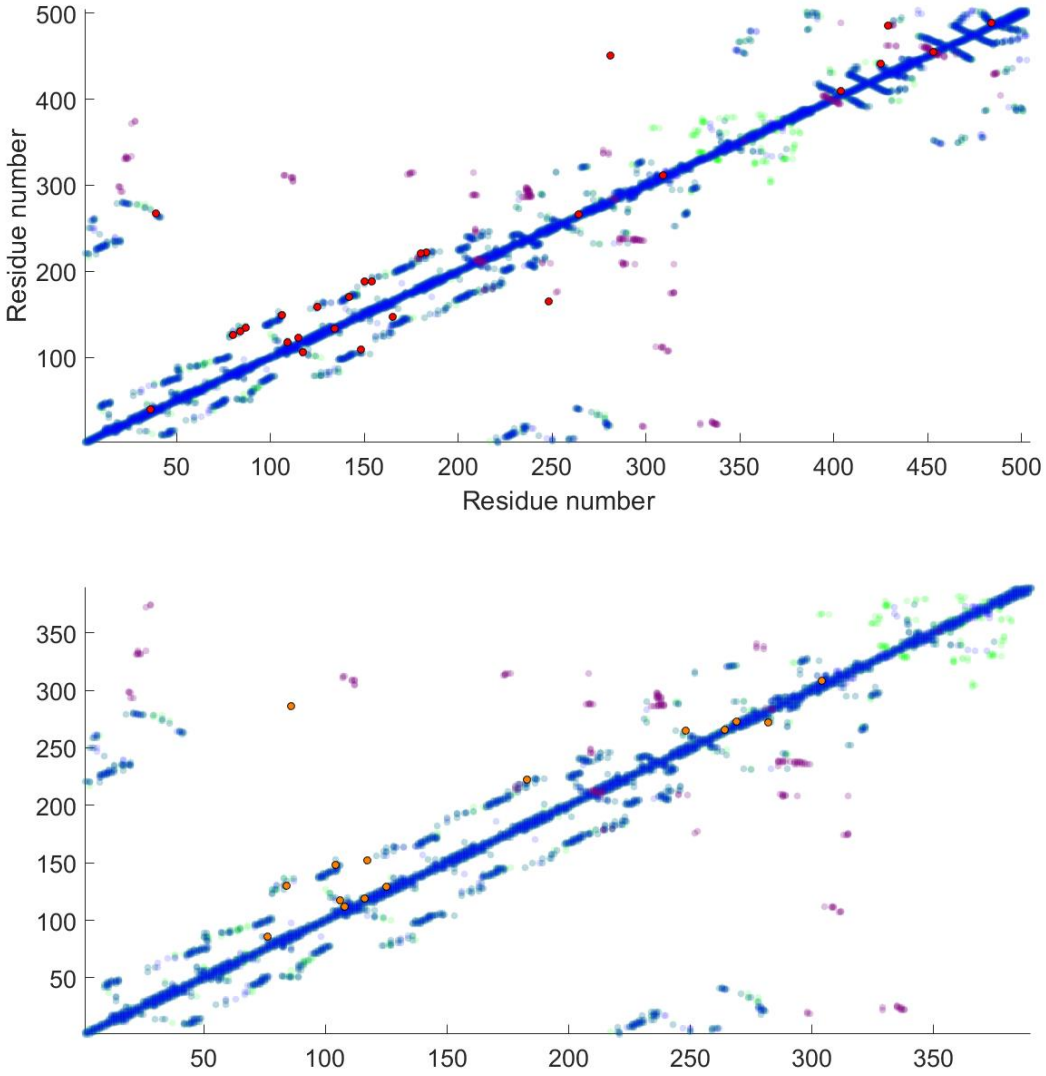


Figure 3.20: Contact maps of the *Nme*IPMS homology model. Chain A is shown in blue, Chain B is shown in green, and cross-chain contacts in purple. The contact maps show the single pair interactions identified by MIp in orange/red. The interactions identified by the RDP MIp analysis are in the top figure, and the interactions identified by the RDA analysis are in the bottom figure.

In the RDP MIp analysis, there is one single pair that does not form close structural contacts, residues 281 and 451 (*Nme*IPMS numbering). Plausibly, these residues may be closer in space in the dynamic protein, or during the catalytic cycle, and thus share mutual information. In the RDA MIp analysis, several single pairs cannot be accounted for by close structural contacts. This is likely because the *Nme*IPMS homology model was used to construct the contact map, and this contains a regulatory domain.

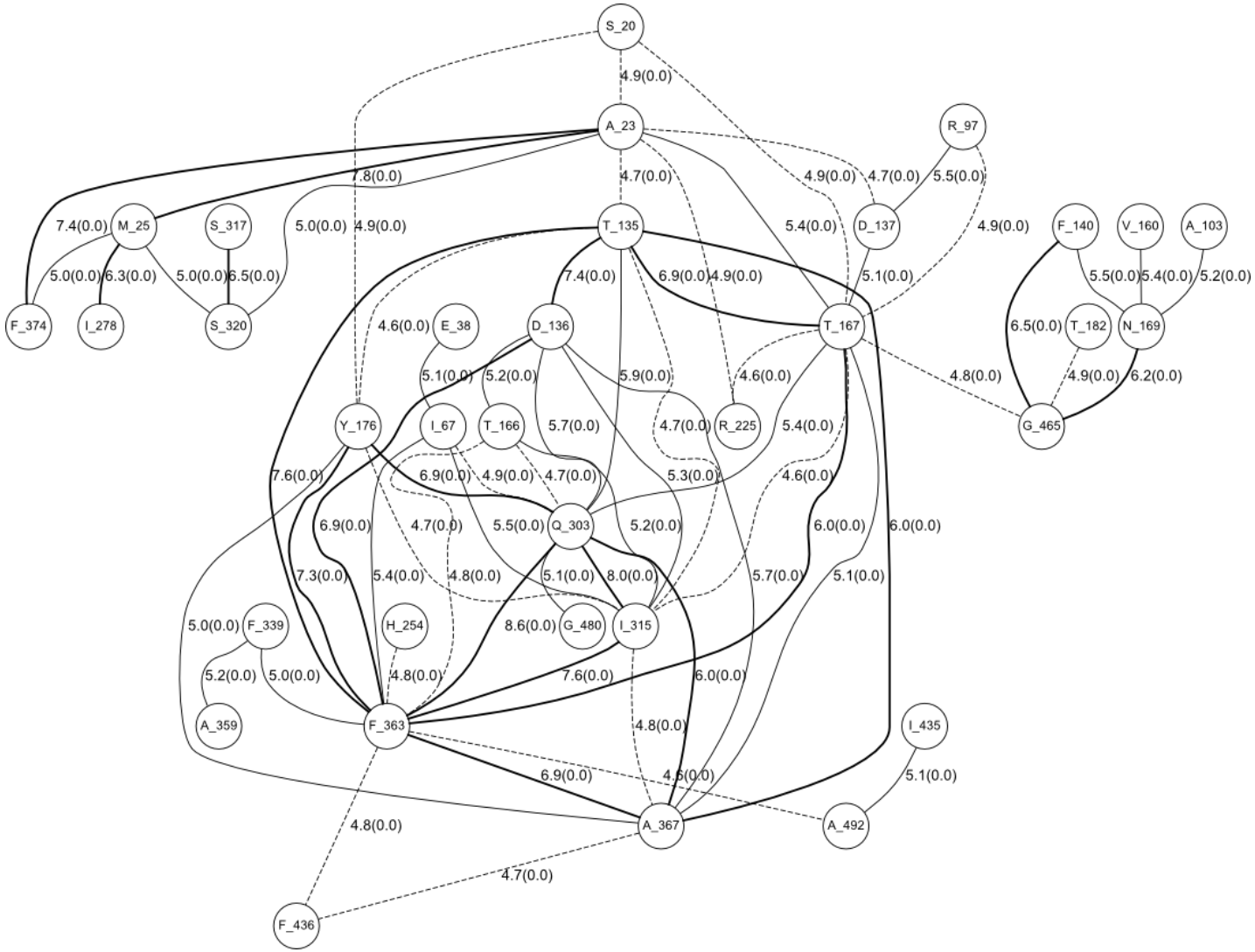


Figure 3.21: Group residues showing mutual information identified using the RDP alignment.

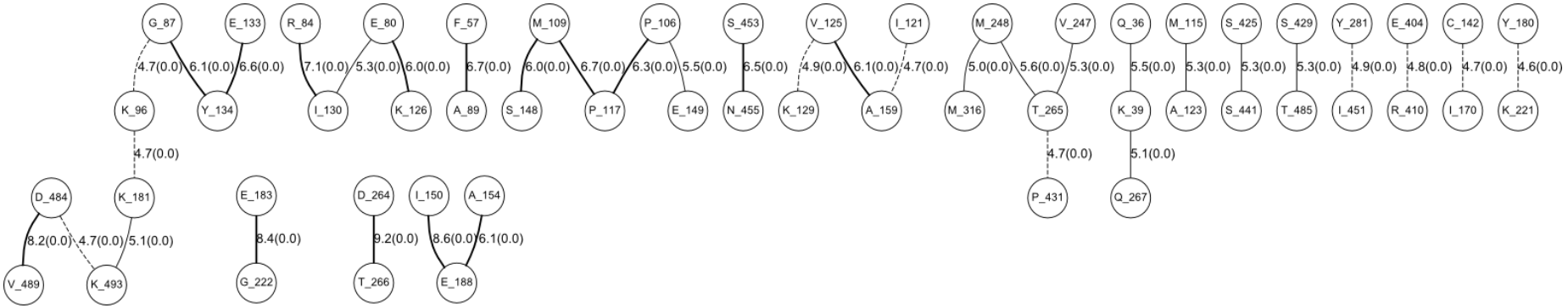


Figure 3.22: Single pair residues showing mutual information in the RDP alignment.

3.4.2.1 RDP sequence population MIp analysis

The so-called 'group' residues (Figure 3.21) are of particular interest as they form a network of mutual information that provides insight into the functioning of the protein.[145] The network can be narrowed down to 'nodes' of residues that show the highest mutual information in relation to each other (Figure 3.23). These residues are from the catalytic domain, subdomain I, and subdomain II. None of the residues appears in the molecular dynamics simulations or docking studies. This could be due to the nature of the sequence alignment, or because the molecular dynamics simulations did not fully account for the natural behaviour of the protein.

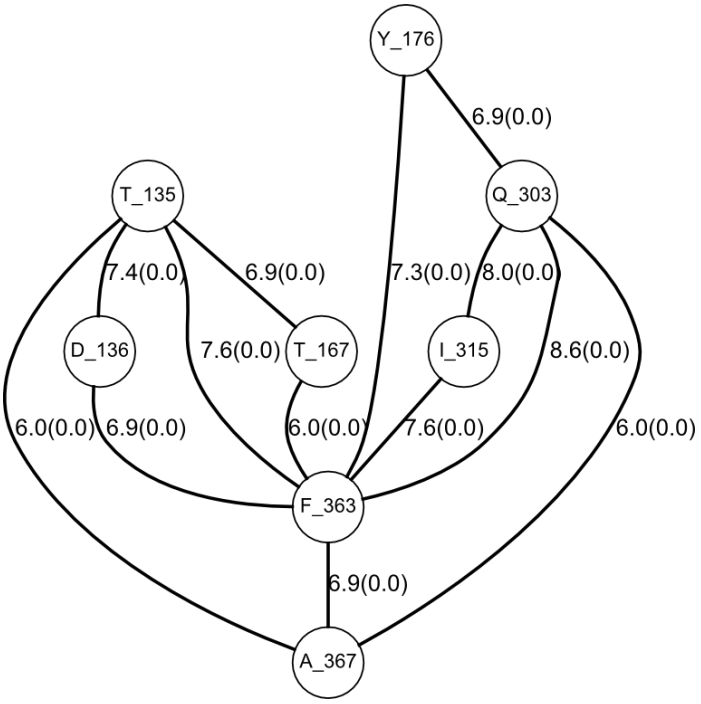


Figure 3.23: The node residues from the RDP MIp analysis that share the highest Z scores.The residue numbering is from *Nme*IPMS.

Interestingly, several of the residues were very close in sequence to highly conserved parts of the protein that are known to be important for catalysis. Residue Tyr176 falls just after a group of highly conserved residues (Pro171-Gly175) that include Thr173, known to be important for catalysis. Residue Ile315 also falls just after a highly conserved residue, Tyr313, which is also important for catalysis. Tyr176 and Ile315 also appear to form a cross-chain hydrophobic interaction in the *Nme*IPMS homology model that may restrain subdomain I. If a structural alignment of the L-leucine bound conformation of *Mtu*IPMS (PDB: 3FIG) against the homology

model of *Nme*IPMS is performed, the two comparable residues, Met257 and Pro412 (*Mtu*IPMS numbering) are also in close proximity in the *Mtu*IPMS L-leucine bound structure (PDB: 3FIG), due to the domain swap that occurs in the homodimeric structures. In the L-leucine absent structure (PDB: 1SR9), the region containing Met257 is disordered. Tyr176 and Ile315 (*Nme*IPMS numbering) do not show significant mutual information in the regulatory domain absent analysis, suggesting that this interaction may be particularly important for those sequences that contain a regulatory domain.

As with the example above, residue Phe363 and Ala367 are also in a region of high conservation, although neither residue themselves is particularly highly conserved. This suggests that the interactions and positions surrounding conserved regions may show mutual information to allow for preservation of the contacts or interactions that those residues within the conserved region. Molecular dynamics simulations suggest that residue Lys364 in *Nme*IPMS forms a hydrogen bond with Glu503 in the presence of AcCoA, suggesting that the interactions adjacent to this conserved residue are important for maintaining its ability to interact with other residues to facilitate AcCoA binding and subsequent catalysis. Phe363 also forms the centre of the node residues, having significant mutual information with 6 other residues. This residue appears to form part of the hydrophobic 'core' of the alpha helical bundle that makes up the subdomains, and thus may be crucial for maintaining proper tertiary structure in that region. The presence and stability of subdomain II has shown to be critical for catalysis, potentially due to subdomain II's role in limiting the conformations formed by the linker region between subdomain I and II that is important for AcCoA binding.[2]

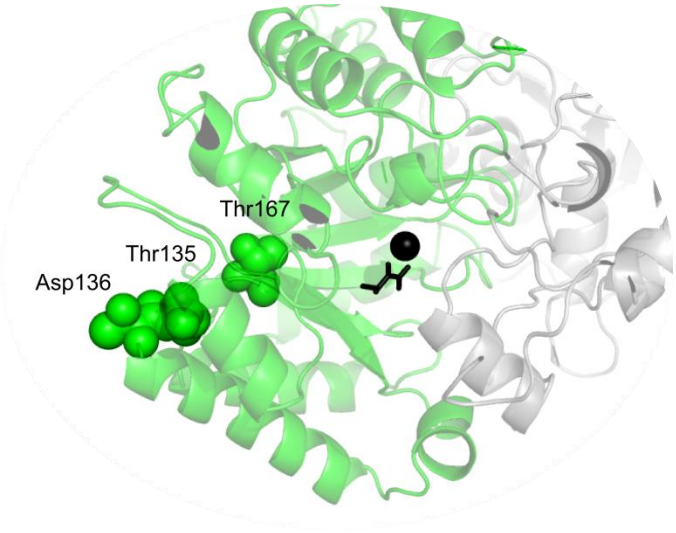


Figure 3.24: The location of residues 135, 136, and 167, in the *Nme*IPMS homology model.KIV and the metal ion are shown in black

Interestingly, the residues Thr135, Asp136, and Thr167, that also form part of the group residues identified by MIp, are on the opposite face of the protein to where subdomain I interacts and to where the active site is located (Figure 3.24). However, they may have a role in aiding the inherent flexibility of the catalytic site to facilitate substrate binding and catalysis, and thus share mutual information with other residues important for facilitation of catalysis.

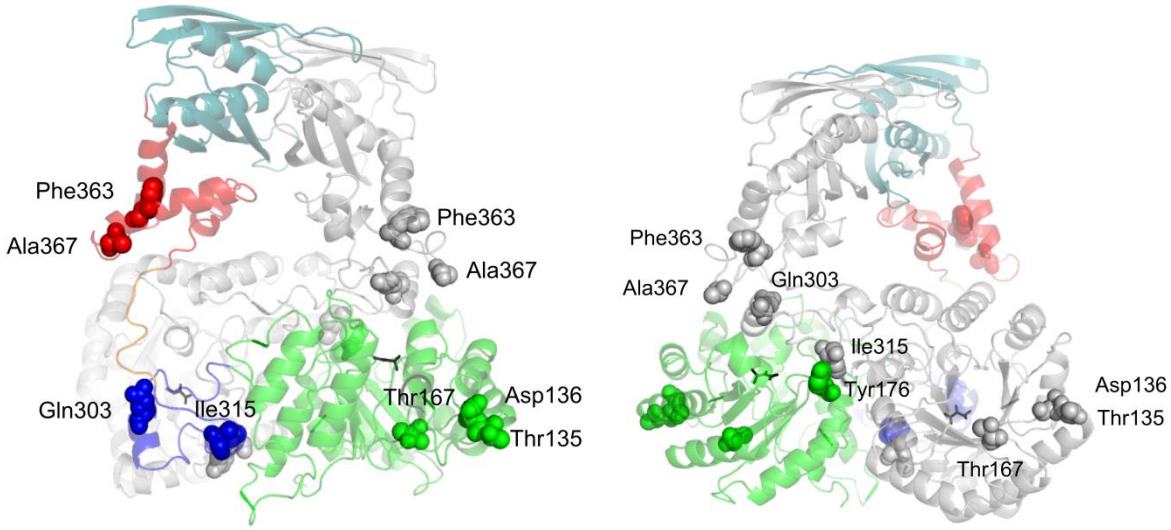


Figure 3.25: The node residues identified by MIp from the RDP alignment mapped onto the homology model of *Nme*IPMS. The homology model is rotated by 180°.

As a whole, the group residues identified in this analysis form a network of interactions from the catalytic domain to the regulatory domain (Figure 3.25), which implies that maintenance of these

interactions has been preserved through evolution. Providing sufficient flexibility to allow for the large-scale dynamic movements predicted to play a role in catalysis yet maintaining enough rigidity and structure to limit the number of conformations the protein can form, particularly in the subdomains, is key in the evolution of this enzyme population.

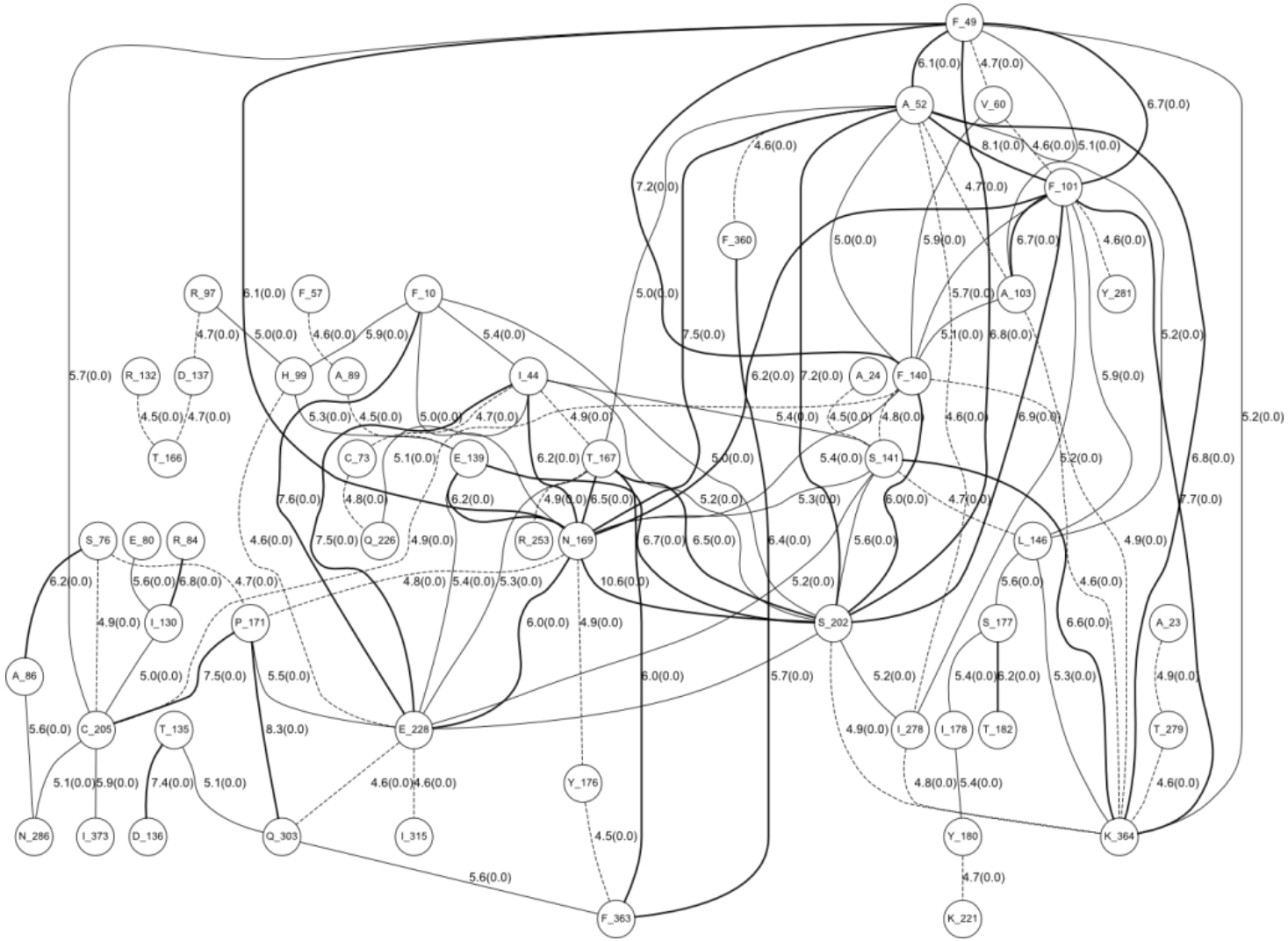


Figure 3.26: The group residues of the RDA MIp analysis.

Figure 3.27: The single pairs from the RDA MIp analysis.

3.4.2.2 RDA sequence population MIp analysis

As with the RDP sequence population, there was a mixture of single pairs sharing mutual information and a group of residues that share mutual information with each other (Figure 3.26). The single pairs are predominantly close in sequence and structure, and thus probably are key for maintaining local tertiary structure or dynamics (Figure 3.27).

The main group of residues that share mutual information (Figure 3.26), and the 'nodes' of this group (Figure 3.28), are different to that of the RDP sequence population. The RDA node residues identified by MIp include a number that are known to be important for substrate selection in the active site, namely residues Phe101, Leu75, Glu139, His99, and Ser141.[69, 72] There is a phylogenetic basis to these residues appearing, as substrate selectivity is phylogenetic, but these residues do show mutual information as knowledge of one residue provides information about other residues important for substrate selectivity in the active site. This mutual information may not have been apparent in the RDP population as Ser141 is also a serine in the *Mja*CMS-like population although not the *Lin*CMS-like population, which may reduce the mutual information in these residues.

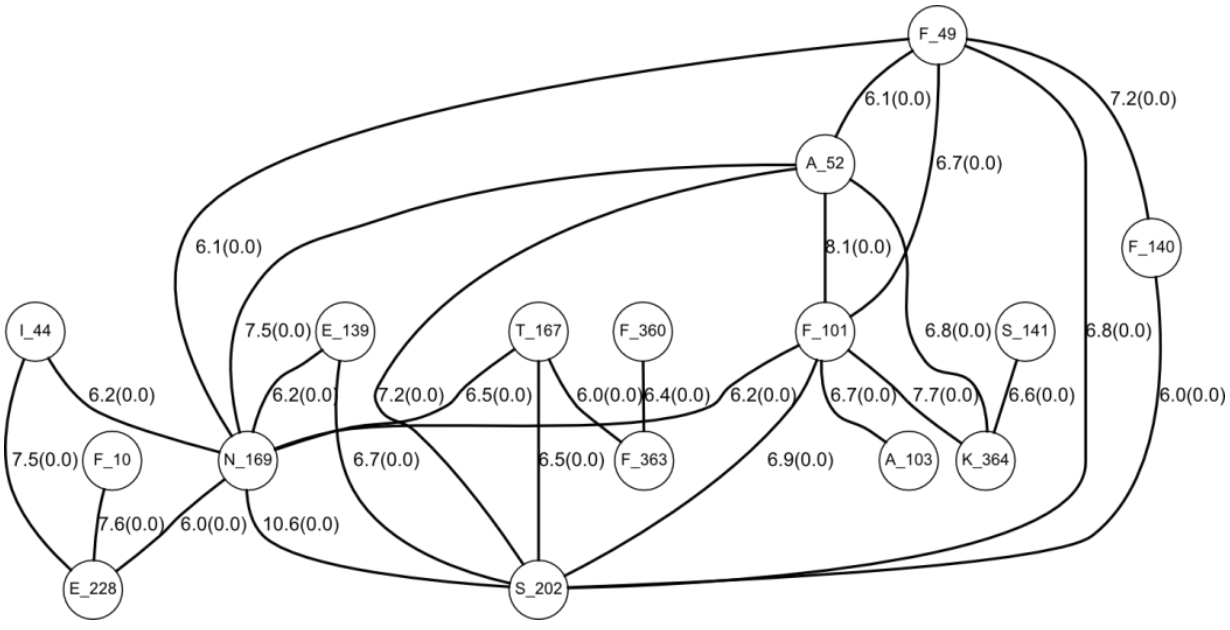


Figure 3.28: The node residues from the MIp analysis of the RDA alignment (*Nme*IPMS numbering).

Residue 364 is also in this group of node residues, and this residue, as discussed above, may form a hydrogen bond with Glu503 in *Nme*IPMS in the absence of leucine. This interaction is broken

in the presence of leucine in the MD simulations. The appearance of residues such as this and Glu298 in the regulatory domain absent, but not regulatory domain present mutual information analysis is somewhat unexpected, but it does suggest that the interactions that are potentially key for allostery, that were identified in the *Nme*IPMS SCA, may also be key for control of the subdomains, and thus may be key for catalysis in the regulatory domain absent population and therefore share mutual information.

Only one residue is found in the 'node' residues of both populations in the MIp analysis, residue Thr167 (*Nme*IPMS numbering). In *Nme*IPMS, this residue is a threonine, while in *Spo*HCS, and *Lin*CMS, it is an arginine. In the structures available of *Spo*HCS, Arg191, the equivalent of Thr167 in *Nme*IPMS, forms different interactions depending on the substrate bound to the active site. In the apo structure of *Spo*HCS, Arg191 forms interactions with Glu222 and Glu161, whereas in the KG-bound structure of *Spo*HCS, the arginine residue forms an interaction with Glu222 and Asp123 (Figure 3.29, top). Arg191 is known to be important for binding of the competitive inhibitor, lysine, and indeed in the lysine-bound structure, Arg191 is forming an interaction with Glu161, allowing Asp123 to flip into the active site to form a hydrogen bond with lysine (Figure 3.29, bottom). In the structures available of *Lin*CMS, the equivalent to Thr167 in this enzyme, Arg173, forms an interaction with Glu9 and Asn142 (*Lin*CMS numbering). The latter residue is located in a position thought to be important for determining active site specificity in these enzymes.[69]

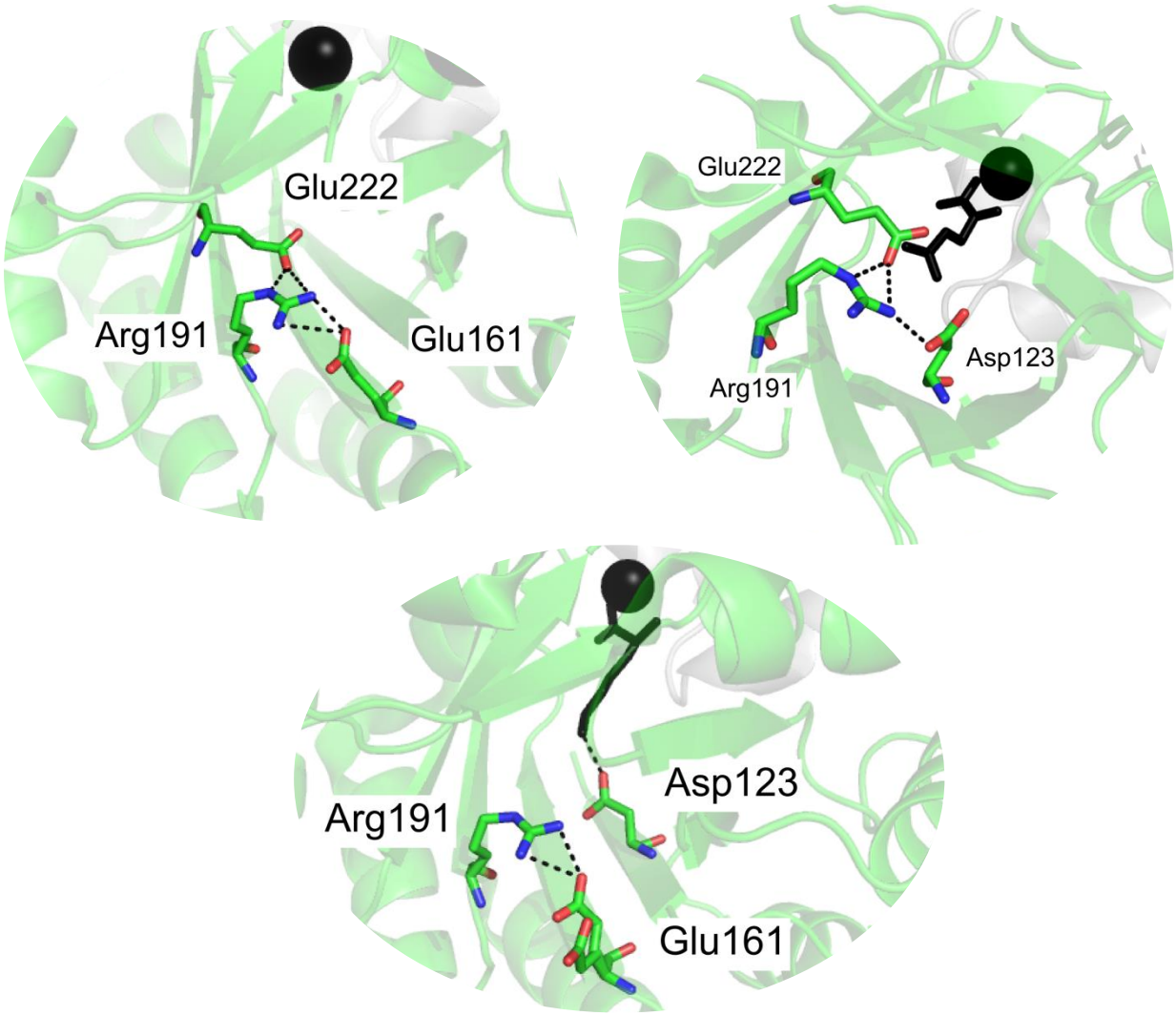


Figure 3.29: The interaction of Arg191 in *Spo*HCS. In the apo structure (PDB: 3IVS, top left), with the substrate ketoglutarate bound (PDB: 3IVT, top right), and with the competitive inhibitor, lysine, bound (PDB: 3MI3, bottom). The metal ion is shown as a black sphere, and the substrate and inhibitor are shown in black.

The comparable residue to Asn142 from *Lin*CMS in *Mtu*IPMS, Glu214, forms a direct interaction with the ketoacid substrate. In the structure of *Tth*HCS with ketoglutarate bound, Arg160, the Thr167 equivalent, forms interactions with Glu193 and Asp92 (Figure 3.30, left). These are equivalent to Ser202 and His99 from *Nme*IPMS, and Glu222 and Asp123 in *Spo*HCS. In the lysine-bound structure of *Tth*HCS (Figure 3.30, right), Arg160 forms interactions with Glu193 and Glu131, allowing, as with *Spo*HCS, the aspartate residue to form interactions with lysine in the active site.

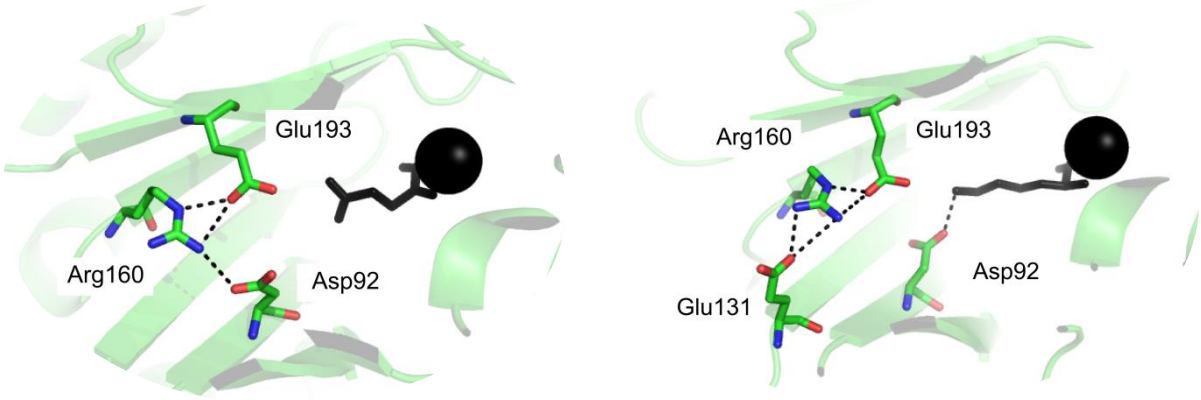


Figure 3.30: The change in hydrogen bond interaction of Arg160 in *Tth*HCS with ketoglutarate (left) and lysine (right) bound.

As the residue in the Thr167 position appears to have a significant influence over the character of the active site, it seems reasonable that it forms substantial mutual information with other residues, especially those involved with substrate or inhibitor binding or selectivity, in both sequence populations under investigation. Additionally, the strong mutual information between Thr167 and Ser202 in the RDA alignment can be rationalised as those residues form interactions in the structures available of the regulatory domain absent population.

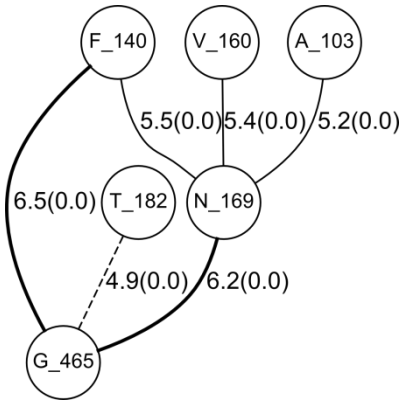


Figure 3.31: The small group of residues in the MIp analysis of the RDP alignment showing residues that share mutual information with Asn169

There are substantial differences in both the 'node' residues in the two populations, as well as the group residues as a whole. In the RDP alignment, Gln303, Ile176, Ile135, Phe363, and Ala367 are all part of the 'node' residues. In the regulatory domain absent population, although Gln303, Phe363, and Thr135 are present in the results of the analysis, they do not form significant MI with each other or other residues as seen in the regulatory domain present population. Conversely, in the regulatory domain absent population, Asn169 is present in the regulatory domain present analysis, but forms a small network alone with Phe140, Gly465, Ala103, and Val160 (Figure 3.31).

Phe140 forms part of the node residues along with Asn169 in the main group in the RDA alignment, suggesting that these residues are important for local interaction in the RDP population.

There is a considerable amount of conservation at position Gln303 in the regulatory domain present population, with approximately 62% having Q at that position. There is much more limited conservation at that position (30% Ala 30% Val) in the regulatory domain absent population. This residue is additionally next to an absolutely conserved histidine (His302 in the *Nme*IPMS sequence). In the KIV bound structure of *Mtu*IPMS, the only full length regulatory domain containing structure available, the equivalent of Gln303, Gln380, forms a sidechain interaction with solvent in both chains. In the leucine bound structure, Gln380 forms an interaction with Gln438, which is equivalent to Lys364 in *Nme*IPMS. Lys364 is found in the node residues of the regulatory domain absent population, and the residue next to it, Phe363, is one of the node residues in the regulatory domain present population.

At position Phe363, there is considerable conservation in the regulatory domain present population with approximately 60% having Phe at that position, and at position 362, nearly ¾ of sequences have either Arg or Lys, while in position 364, 86% have a lysine. This regulatory domain containing population, as mentioned above, does not include *Mtu*IPMS-like sequences due to problems with sequence alignments. In the regulatory domain absent population, the majority of sequences have V or I at position 363, whereas position 364 has approximately 80% of sequences having Arg or Lys at that position. There is almost no sequence conservation at position 362. This suggests that position 363 may be key for control of the subdomains in the regulatory domain containing population, but not in the regulatory domain absent population.

The comparison of the two populations demonstrate that although there are similarities, namely in the catalytic domain, there are considerable differences, especially in the subdomains, as to which positions share substantial mutual information. Interestingly, although there is a major difference in structure between the RDP and RDA populations, and the relative flexibility or inflexibility, of the subdomains appears to be key for the catalytic function in both populations, there is a limited number of positions that share substantial mutual information in either population, although there are considerable differences between the two populations. However, both populations produce a group of positions that share mutual information that form a network from the catalytic barrel through the subdomains, and to the regulatory domain in the population that has one. Additionally, a limited number of residues in either population is identified by the molecular dynamics simulations. This may indicate that there are considerable, potentially phylogenetic, differences in dynamics between groups of sequences, and that the interactions

involved in maintaining dynamics may not share substantial mutual information with other residues, i.e. they evolve independently. However, the presence or absence of positions in either population, particularly in subdomain II, sharing mutual information demonstrates that there are considerable differences between the two structural populations, and these can be used to hypothesise about the roles of these residues in maintaining the catalytic function of the enzyme in the presence and absence of a regulatory domain.

3.5 Comparison of MIp and SCA results

In this study, both SCA and MIp used the same multiple sequence alignment for each sequence population and there were a number of similarities in the results. The results of MIp are displayed as a network diagram, with a mixture of single pairs of residues sharing mutual information as well as a group of networked residues that share mutual information amongst themselves, while the results of SCA are a single sector of residues that show statistical coupling. Of the residues in IC1 from the RDP SCA, several were also found in the group residues from the MIp analysis for the same alignment (Figure 3.32). These residues are of particular interest as they comprise of residues from the catalytic domain and the subdomains (Figure 3.32). As these residues are in both IC1 and the MIp analysis, it suggests that the network of these residues show coevolution and may be important for the control of the subdomains to facilitate catalysis.

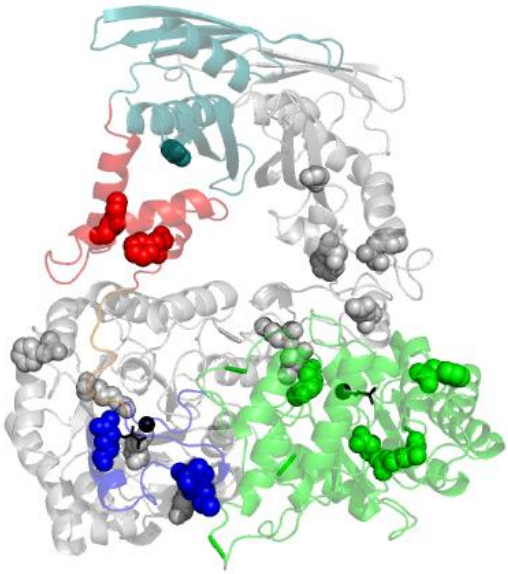


Figure 3.32: Residues identified by both covariance analyses methods performed on the RDP alignment mapped onto the *Nme*IPMS homology model.

Residues that were identified by both covariance methods (*Nme*IPMS numbering)
57
97
169
176
303
315
316
339
363
480

Table 3.1: Residues identified by both covariance analyses performed on the RDP alignment

Aside from one residue, that also appears in IC3, no residues from IC2 are present in the MIp analysis. As discussed above, this IC includes residues that show statistical coupling due to the *Lin*CMS-like CMS group showing substantial phylogenetic distance from the other sequences in

the population. As MIp is phylogeny independent, it stands to reason that this group is not represented in the MIp analysis. IC3 is comprised of a number of residues that also appear in the MIp analysis. These residues are predominantly from the single pairs in the MIp analysis, suggesting that these couple independently of the group residues.

As with the RDP alignment, there are several residues in the RDA alignment that show both statistical coupling and mutual information (Figure 3.33). As with IC1, these residues show a network of residues that are present in the catalytic and subdomain II, suggesting that these residues are important for the facilitation of catalysis. The only residue that shows mutual information and statistical coupling in both populations is Asn169, suggesting that this residue is crucial for catalysis in both populations.

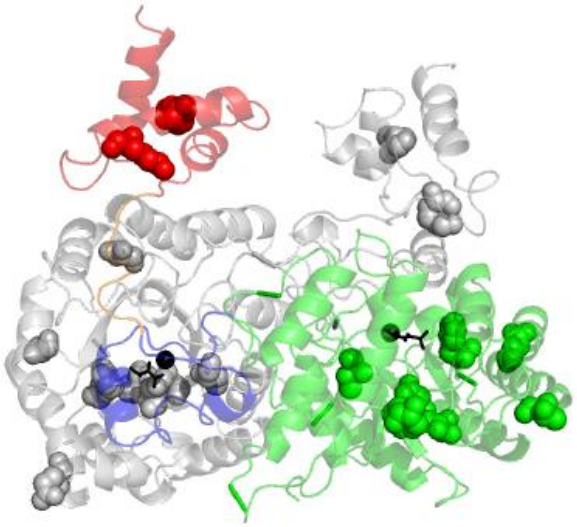


Figure 3.33: Residues from the RDA alignment that show both mutual information and statistical coupling (spheres) mapped onto the *Nme*IPMS homology model truncated at residue 394.

Residues that were identified by both covariance methods (*Nme*IPMS numbering)
68
101
119
135
167
169
202
205
360
364

Table 3.2: Residues identified by both covariance analyses performed on the RDA alignment

Although some residues are similar in the two populations, there are also substantial differences, especially with the residues that show lower levels of either statistical coupling or mutual information, demonstrating the difference between the two algorithms. The purpose of the conservation weighting in the SCA is to reduce the amount of noise in the alignment by weighting by conserved residues, with the rationale being that changes in conserved residues are more likely to have an influence on the functionality of the protein. There are substantially fewer residues showing statistical coupling than those showing mutual information, suggesting that the weighting does highlight more important residues, although whether up-weighting conserved residues means

that residues that are not conserved but are coupled are missed by the SCA remains to be seen. These analyses suggest that both algorithms could be used to identify coevolved residues, although more work would be needed to assess which provides the better information for these proteins. This could be achieved by site-directed mutation followed by kinetic analysis to assess the validity of the networks, or specific residues in the network, in relation to their role in catalysis.

Although most of the network was not identified by both techniques, in both populations, a group of residues were identified by both techniques. These residues that demonstrate both statistical coupling and mutual information appear to be a network that may be involved in conveying catalysis. The networks in each population, aside from residue 167, are different, suggesting a mechanism by which the separate structural groups control the subdomains in the presence and absence of the regulatory domain.

3.6 Overall discussion

Probing sequence information for coevolution information has been a technique that has undergone many changes over the years. There is some debate about the validity of particular analyses and algorithms, but there is debate also about whether coevolution can even be detected by analysis of multiple sequence alignments. Talavera et al.[146] posits the "coevolution paradox" that states that coevolution would need to be sufficiently strong to cause coordinated changes, and this strength would reduce the evolutionary rate to such a level that they are unlikely to happen. However, Avila-Herrera et al.[147] suggests that a neutral mutation or one mildly harmful to fitness could occur, and alter the fitness landscape in such a way to increase the likelihood of a compensatory mutation occurring to maintain structure or function. Both mutual information and statistical coupling analysis have been used in numerous studies where experimental evidence that these algorithms and the subsequent analysis of the results can provide information about structure-function relationships in a protein that other analyses cannot.[120-122, 124, 129, 143, 145, 148, 149]

As discussed above, the conservation weighting function used in SCA can bias the results of the algorithm particularly when there are very different phylogenetic groups as there were in this analysis. This presents an obvious problem when searching for coevolved, yet not coevolved and conserved, residues and one that could be worked around by changing the weighting function, as discussed by Colwell et al.[139], or by removing it entirely by using an earlier version of the SCA programmes.

Mutual information, in its MIp form, was therefore employed to assess whether there is substantial coevolution that was not tied to phylogeny. However, MIp is not without its flaws. Some of these, such as sequence alignment errors, are fundamental to any of these analyses that are dependent on an accurate sequence alignment to represent evolution. Additionally, indirect coevolution in mutual information analyses, where a pair of residues appears to have higher mutual information due to their high mutual information with another partner, can skew analyses by making long-distance contacts appear to have high MI, when indeed it is their interaction with closer structural partners, also with high MI, that is actually correct. To combat this issue, a Bayesian network model has been produced to attempt to sort direct from indirect dependencies.[150] Additionally, a new approach called Bayesian Partitioning with Pattern Selection has been pioneered, which may present a way forward to improved techniques for coevolution analysis.[151]

In these analyses, the goal was to identify residues that contribute to the overall function of the protein, particularly with regards to the very flexible subdomains that have a crucial role in catalysis and can bear the burden of an additional structural element present in some proteins but not others. The two analyses produced similar results with differences that can be accounted for by the difference in the way the output of the algorithm is weighted with regards to conservation. However, both analyses produced results that correspond well with information already known about the protein, and a considerable number of residues is found in both MIp and SCA analyses. These analyses identified some residues that could be of considerable interest going forward that would not have been otherwise apparent by analysis of the alignments in light of conservation. Additionally, these results can later be mined for information that correlates with other knowledge about the proteins, such as point mutations and subsequent characterisation, to continue to build a picture of these dynamic proteins.

Chapter 4: A truncated form of *Nme*IPMS

4.1 Introduction

The role of the regulatory domain in the IPMS and CMS proteins is clear: the allosteric inhibitor is bound by this domain and the allosteric signal is somehow transmitted to the distant active site. However, the proposed evolutionary trajectory suggests that the ancestral IPMS did not have a regulatory domain. It is not clear, therefore, how the very flexible subdomains, whose motion is critical for catalysis, can compensate for the comparative burden of the regulatory domain in some of the extant proteins. The information obtained in Chapter 3 about the difference in coupled residues in the subdomains in the presence or absence of a regulatory domain highlights some similarities and considerable differences between the two populations of proteins. However, the relative flexibility and stability of the subdomains has been shown to be critically important in both structural populations. It has, however, been shown that IPMS and CMS proteins that have evolved to bear the burden of the regulatory domain can provide catalysis in the absence of a regulatory domain.[2, 108]

It was reported previously that *Nme*IPMS could not perform catalysis without a C-terminal regulatory domain.[1, 64] However, it has since been determined that the truncation used in these analyses was also missing a considerable part of subdomain II as well as the regulatory domain, as the location of the truncation was at residue Glu365, and more recent homology model construction and sequence alignments have suggested that subdomain II of *Nme*IPMS encompasses residues up to residue Lys395. To explore further how the subdomains facilitate catalysis in the presence and absence of the regulatory domain, a truncation of *Nme*IPMS was made at a different location to that previously reported to encompass all of subdomain II.

This truncated protein provides insight into the evolution of these proteins, as the catalytic module of the catalytic domain and subdomains exist in contemporary proteins in both allosterically regulated, that also have a C-terminal regulatory domain, and unregulated forms. Additionally, this truncated and active enzyme permits the study of the subdomains in the absence of their role in allosteric regulation and provides additional insight into the importance of dynamics in the catalytic cycle.

4.1.1 Truncation of *Nme*IPMS by site-directed mutagenesis

The location of the point at which to truncate *Nme*IPMS to ensure all of subdomain II was included within the truncated protein was determined using multiple sequence alignments, pairwise sequence alignments, and structural alignments with known structures (Figure 4.1). The precise residue that was changed to a stop codon was selected as it was in a short region that did not show substantial conservation in the MSA, suggesting that that particular region was not of particular importance for allostery or catalysis. The residue selected, Lys395, is also located after the end of subdomain II in the MSA with a short form IPMS (*Lbi*IPMS short) and the *Spo*HCS, suggesting that subdomain II in the truncated *Nme*IPMS would be intact.

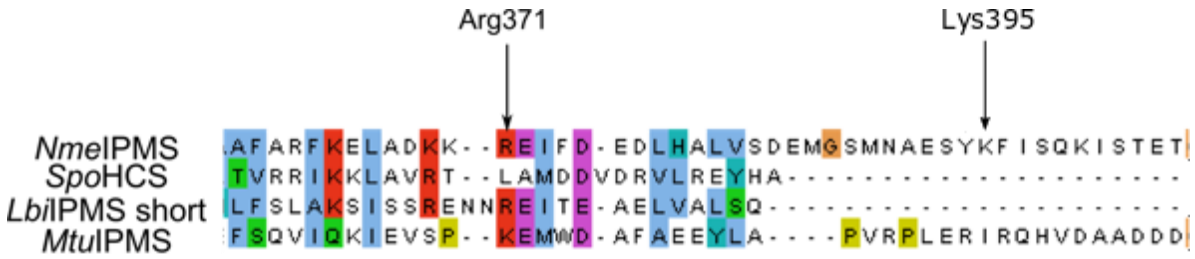


Figure 4.1: A MSA showing the position of the Lys395 residue in *Nme*IPMS. The MSA was a structural alignment constructed using Promals3D and a structure of the short form of *Lbi*IPMS (PDB: 4OV4), *Spo*HCS (PDB: 3VIS), and *Mtu*IPMS as well as the sequence of *Nme*IPMS. The position of Arg371 in subdomain II is also highlighted. The residue numbers used are *Nme*IPMS numbering.

The truncation of *Nme*IPMS was made by site-directed mutagenesis (SDM) to change the codon coding for Lys395 (AAA) to a stop codon (TAA), thus truncating the protein at residue Tyr394 (Figure 4.2). This protein will be discussed as *Nme*IPMS K395Term. The construct used for SDM was the wild type *Nme leuA* gene cloned into pET151 by TOPO cloning as described and performed by Huisman[64]. The presence of the altered codon was assessed by DNA sequencing and SDS-PAGE (Figure 4.3).

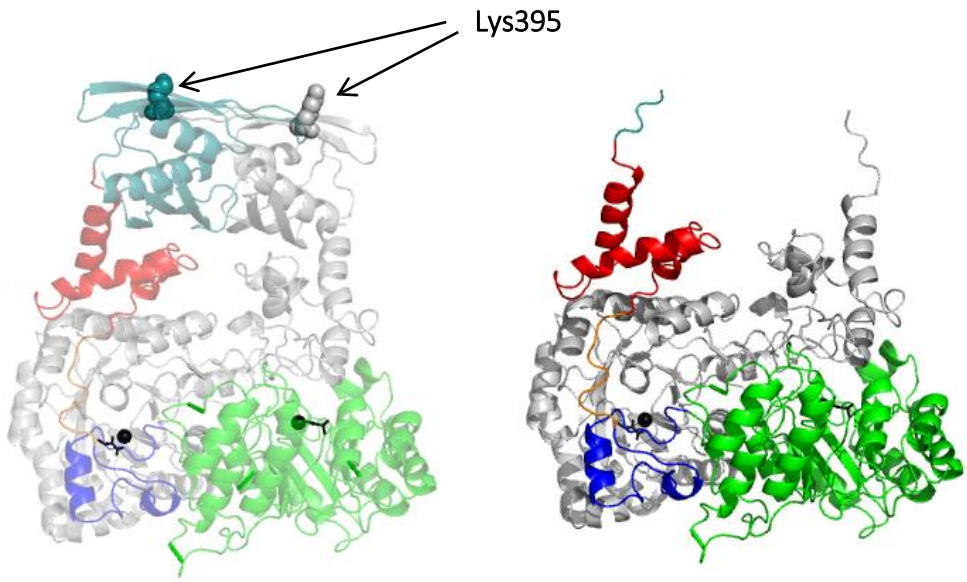


Figure 4.2: The homology model of *Nme*IPMS with residue Lys395 highlighted as spheres (left), and the same model with the residues beyond the point at which the truncation was made removed.

4.1.2 Purification of *Nme*IPMS K395Term

Purification was performed using the N-terminal polyhistidine (His_6) tag, and immobilised metal affinity chromatography. The stability of the truncated protein appeared to be compromised, as activity was not maintained through the purification procedure if the His_6 tag was removed. This instability does not appear to be related to the presence or absence of the His_6 tag, as the small amount of un-tagged protein that was obtained was kinetically similar to tagged protein, but due to the excess handling of the protein through multiple purification steps, that caused soluble aggregates to form. Therefore, the His_6 tag was not removed, and the protein underwent a gel filtration step to produce acceptably pure protein for further studies (Figure 4.3). When the truncated protein is referred to in this thesis, it refers to the His_6 tagged protein.

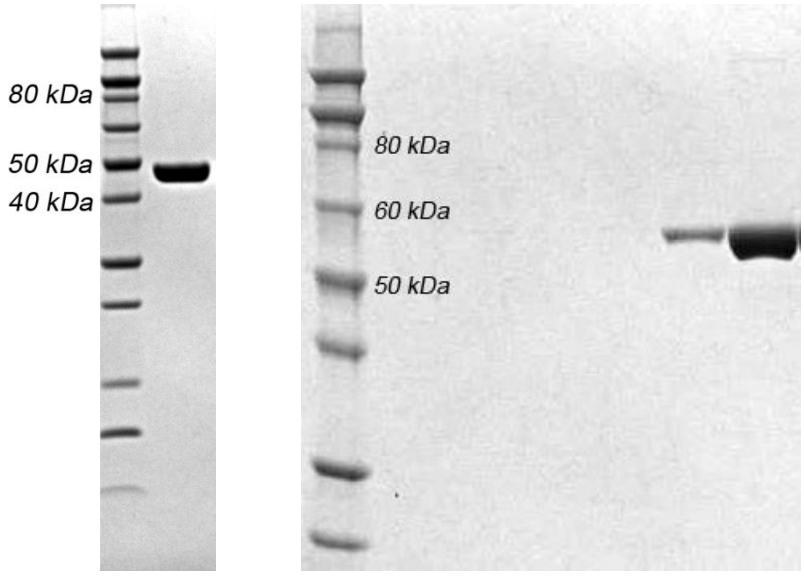


Figure 4.3: Purification of *Nme*IPMS K395Term (left) and *Nme*IPMS wild type, demonstrating the difference in size

4.1.3 Differential scanning fluorimetry

Differential scanning fluorimetry (DSF) was used to explore the stability of the protein. Zhang et al.[2] noted that truncations of *Lb*IPMS2, *Lb*IPMS2-R376 (equivalent to *Nme*IPMS Arg371) and *Lb*IPMS2-S387 (equivalent to *Nme*IPMS Asp384), were unstable and suggested this was due to exposure of the hydrophobic core of subdomain II. Thermal shift assays were utilised to explore the thermal stability of the *Nme*IPMS truncation compared to the wild type protein.

Thermal shift assays using the wild type protein compare well with previous results (Figure 4.4).[64] A single peak was seen in all thermal shift assays performed using *Nme*IPMS and *Nme*IPMS Lys395Term. There is a moderate increase in temperature when the substrate KIV is added to the wild-type protein, but a much larger increase was observed when leucine is added, suggesting that presence of 1 mM leucine stabilises the protein. The truncated *Nme*IPMS has a similar T_m to that of the wild type protein with no ligand added, suggesting that truncation of the regulatory domain has not substantially decreased thermal stability. However, unlike the wild type protein, an increase in thermal stability was not observed upon addition of KIV, suggesting that this substrate may not have the same stabilising effect on the truncated protein as it does on the wild type protein. Additionally, there is no increase in thermal stability upon addition of 1 mM L-leucine, demonstrating that the thermal stability provided to the wild type protein upon binding the allosteric inhibitor is not observed in the truncated protein, suggesting that it is not stabilised by L-leucine in the absence of a regulatory domain.

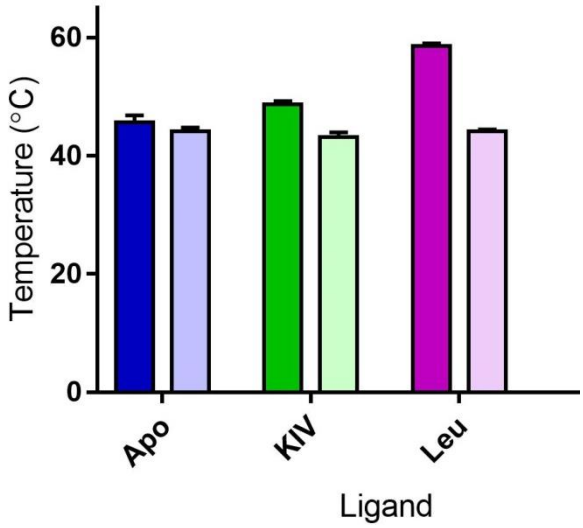


Figure 4.4: Thermal melt temperatures determined by DSF for *Nme*IPMS (darker shades) and *Nme*IPMS K395Term (lighter shades) in the presence of different ligands.

4.1.4 Analytical SEC of *Nme*IPMS K395Term

Analytical size exclusion chromatography (analytical SEC) was used to explore the oligomeric structure of *Nme*IPMS K395Term. Previous work has demonstrated that wild type *Nme*IPMS is dimeric in the presence and absence of L-leucine.[64] *Mtu*IPMS also is dimeric in solution in the presence and absence of L-leucine.[74] In a buffer containing 10 mM Tris, pH 8, *Nme*IPMS K395Term appeared to form two distinct species (Figure 4.5, left). Both species contained only the protein of interest, and both species showed similar specific activity when tested. When the salt concentration of the buffer used for analytical SEC increased to 300 mM KCl, only one, albeit broad, peak was present (Figure 4.5, right).

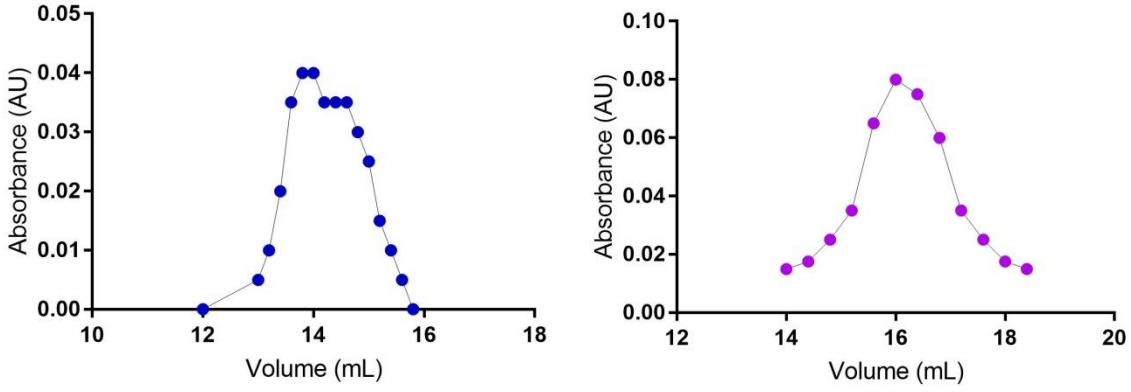


Figure 4.5: Chromatograms of *Nme*IPMS K395Term during analytical SEC. (Left) Analytical SEC of *Nme*IPMS K395Term run in 10 mM Tris, pH 8, 50 mM NaCl. (Right) Analytical SEC run in 10 mM Tris, pH 8, 300 mM KCl. Note the volumes for the low salt buffer (left, blue) are inaccurate due to a computer recording issue. Analytical SEC was performed using a Superdex™ 200 10/300 GL column (GE Healthcare).

This analysis suggests that the presence of salt has a major impact on the oligomeric state of the truncated protein. Wulandari et al.[94] noted that two peaks, thought to be monomer and dimer, of *Tth*HCS were observed when analytical SEC was used to investigate the oligomeric state of *Tth*HCS. Multiple peaks of HCS from *Saccharomyces cerevisiae* were also observed when analytical SEC were performed.[152] Full length *Mtu*IPMS was reported to be only dimeric in solution.[74] The multiple peaks seen in analytical SEC of proteins without a regulatory domain that are not seen in proteins with a regulatory domain suggests that oligomeric stability may be affected by the removal of a regulatory domain even in proteins that have evolved further in the absence of a regulatory domain. Specifically, for *Nme*IPMS K395Term, the analytical SEC results suggest that the protein's oligomeric state is not stable under some conditions, and this may contribute to the protein instability noted above.

Under conditions where a single species was observable, the truncated protein eluted from the column at a volume that was consistent with a dimer. The molecular mass of the species was calculated to be 101 kDa, whereas the estimated molecular weight of a dimer was calculated to be 93 kDa. This result suggests that truncation of the regulatory domain has not altered the oligomeric state in solution in the presence of higher salt concentrations.

4.1.5 Kinetic characterisation of *Nme*IPMS K395Term

4.1.5.1 The kinetics of His_6-tagged *Nme*IPMS

Due to the stability issues with the truncated form of *Nme*IPMS, the N-terminal His_6 tag and TEV protease site, were retained for both the wild type protein and the truncated protein.

Davies[132] reported that the un-tagged *Nme*IPMS wild type protein had a k_{cat} of 8.9 ± 0.1 s^{-1} and that the tagged wild-type *Nme*IPMS protein had a k_{cat} of 1.53 s^{-1}, 6-fold lower than that of the un-tagged protein. The k_{cat} of tagged *Nme*IPMS determined in this study was 7.2 ± 0.1 s^{-1} suggesting that there is a substantial change in k_{cat} on a purification to purification basis and that the N-terminal His_6 tag may have an impact on the kinetics of the protein. In this study, the N-terminal His_6 tag was retained on all proteins, including the alanine mutants discussed in Chapter 4.1.9, and the truncated form of the protein discussed in Chapter 4.1.5.2, and when the proteins are referred to, the protein contains a His_6 tag unless otherwise specified.

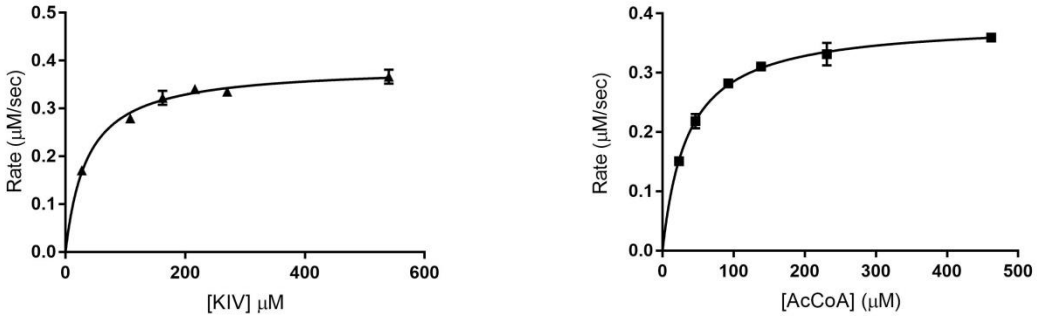


Figure 4.6: Plots of kinetic data of His_6-tagged *Nme*IPMS wild typeshowing the variation in initial rate with change in KIV concentration (left) and AcCoA concentration (right). During these measurements, the other substrate was held at a saturating concentration of 250 μM. The apparent K_m was obtained using the Michaelis-Menten fit from GraphPad Prism 7.00.

Table 4.1: Kinetic parameters of the un-tagged *Nme*IPMS and the His_6-tagged *Nme*IPMS. *The kinetic data for the un-tagged *Nme*IPMS was obtained from Huisman[64].

	K_m (KIV)	K_m (AcCoA)	k_{cat} (s^{-1})
*Nme*IPMS un-tagged*	30 ± 2	35 ± 3	12.8 ± 0.3
*Nme*IPMS His_6-tagged	36 ± 3	35 ± 3	7.2 ± 0.1

4.1.5.2 Michaelis-Menten kinetics of the His_6-tagged truncated *Nme*IPMS

Michaelis-Menten kinetic data of the His_6-tagged truncated *Nme*IPMS were also obtained (Table 4.2). The truncated variant had a similar K_m for KIV to that of the wild type protein, suggesting that loss of the regulatory domain had not adversely impacted the residues involved in KIV interaction. Removal of the regulatory domain and part of subdomain II, although it led to inactive protein, did not adversely affect KIV binding in *Mtu*IPMS or *Nme*IPMS as determined by ITC.[64] Although K_m does not correlate directly with substrate binding, the lack of impact on KIV binding in the inactive truncation suggests that removal of the regulatory domain and subdomain II does not adversely affect the ability of the protein to interact with this substrate.

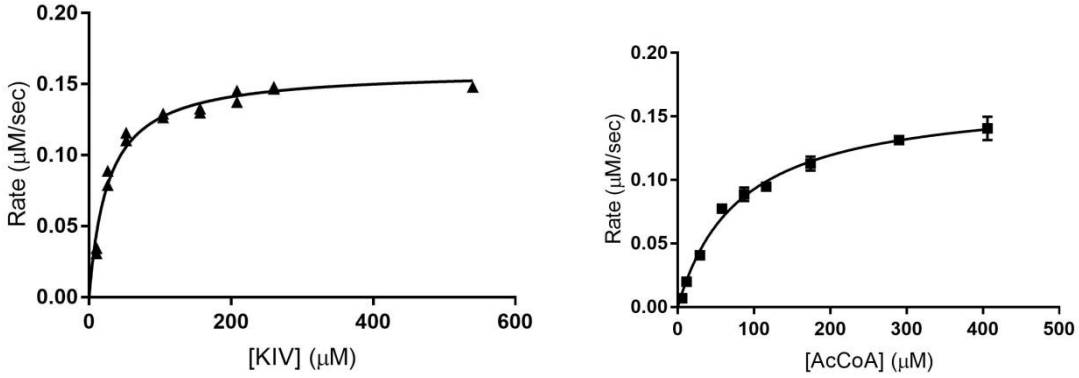


Figure 4.7: Left: Plot showing the change in initial rate of *Nme*IPMS K395Term as the concentration of KIV increases.[AcCoA] was held at a saturating concentration of 250 μM. Right: Plot showing the change in initial rates of *Nme*IPMS K395Term as the concentration of AcCoA increases. [KIV] was held at a saturating concentration of 250 μM.

Unlike the K_m for KIV, the apparent K_m for AcCoA increased two-fold compared to that of the wild type protein, from 35 μM to 80 μM. An increase in the apparent K_m for AcCoA but not one for the other substrate was also seen in a truncated variant of *Lb*IPMS2, where a truncation was made into subdomain II and the protein retained the ability to catalyse the condensation of ketobutyrate and AcCoA although not the natural substrate.[2]

As *Nme*IPMS has evolved to bear the burden of a regulatory domain, the dynamics of the subdomains may have been restrained to compensate for this. Removal of the regulatory domain could increase in the number of conformations the subdomains can form, as, following the truncation of the regulatory domain, the subdomains are only constrained at one end as opposed to two as in the full-length protein.

As the subdomains are known to be important for the recruitment of AcCoA, there may only be certain conformations in which this is possible. An increase in the conformations sampled by the subdomains may lead to an increase in AcCoA K_m, as there is less chance of the subdomains being in an AcCoA-interaction compatible conformation.

Table 4.2: Kinetic parameters of *Nme*IPMS and *Nme*IPMS K395Term. Both of these proteins still contain a His_6 tag. K_m^{app} indicates an apparent K_m. The invariant substrate was held at 250 μM.

	K_m^{app} (KIV)	K_m^{app} (AcCoA)	k_{cat} (s^{-1})
*Nme*IPMS	36 ± 3	35 ± 3	7.2 ± 0.1
*Nme*IPMS K395Term	30 ± 3	80 ± 7	4.1 ± 0.1

As mentioned above, the k_{cat} can be dependent on the enzyme preparation used for the kinetic measurements, although the K_m for either substrate did not change with different purifications. Determination of the k_{cat} requires a known enzyme concentration but the method of determining protein concentration used in this experiment, that of absorbance at 280 nm, does not differentiate between active and inactive enzyme. It is plausible that the enzyme preparation of *Nme*IPMS K395Term and *Nme*IPMS wild-type included some amount of inactive enzyme, and this amount differed between preparations, causing the observed change in k_{cat}. A method of determining the amount of active protein within a particular protein preparation would be advantageous in determining k_{cat} values particularly for this protein.

4.1.5.3 Inhibition of wild type *Nme*IPMS and truncated *Nme*IPMS by L-leucine

The IC_{50} of *Nme*IPMS wild type for L-leucine was obtained (Figure 4.8, Table 4.4). This was comparable to prior results, suggesting that the presence of the His_6 tag does not affect leucine inhibition.[130] The absence of the regulatory domain in the truncated form, as indicted by site-directed mutagenesis and SDS-PAGE, means that no regulation by L-leucine was observed with the truncated protein (Figure 4.8).

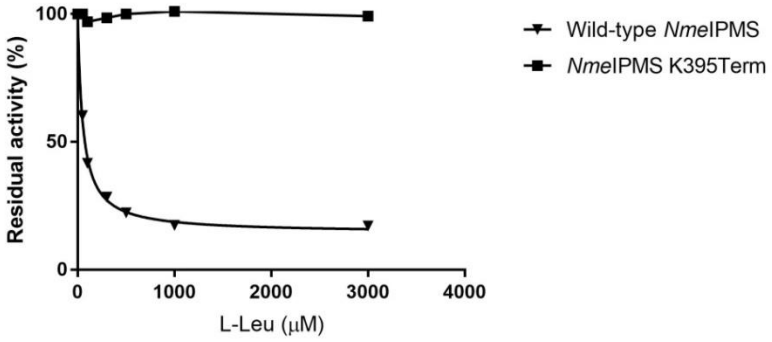


Figure 4.8: The inhibition of wild-type *Nme*IPMS and *Nme*IPMS K395Term to L-leucine. The concentration of both substrates was held at 250 μM for both proteins as the concentration of L-leucine was increased.

4.1.5.4 Cooperativity in *Nme*IPMS

The potential cooperativity of *Nme*IPMS and the truncation were assessed using Hill plots (Figure 4.9). The Hill coefficient, h, denotes negative cooperativity if it is less than one, or positive cooperativity if it is more than one. Neither KIV nor AcCoA showed cooperativity in the wild type protein although cooperativity has been reported in some examples of IPMS and related enzymes.[2, 75, 76] Substrate cooperativity has previously not been reported in *Nme*IPMS.[64]

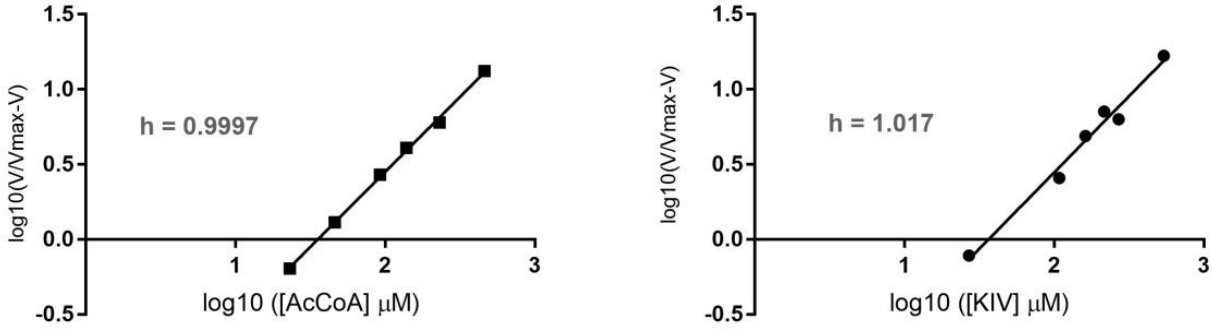


Figure 4.9: Hill plots of wild type *Nme*IPMS demonstrating the absence of cooperativity in substrate binding. h = Hill coefficient.

Cooperativity was not seen in the truncated form of *Nme*IPMS with respect to either for KIV or AcCoA (Figure 4.10). This suggests that either the two active sites are truly independent, or that the cooperativity is being masked somehow. The positioning of the subdomains, particularly where subdomain I forms part of the active site of the opposite chain, suggests that cooperativity between the two chains may be important for catalytic function.

A recent study using statistical mechanics models suggests that the absence of a sigmoidal steady-state kinetics curve does not mean the absence of cooperativity.[153] There can be weak cooperativity that does not result in a sigmoidal curve. Additionally, Moffitt et al.[154] reconciles the lack of a sigmoidal binding curve when ATP binds a homomeric ring ATPase by suggesting that this arises because the binding events are separated by an irreversible transition. Plausibly, there may be cooperativity between the two active sites, it is simply not apparent using steady-state kinetics.

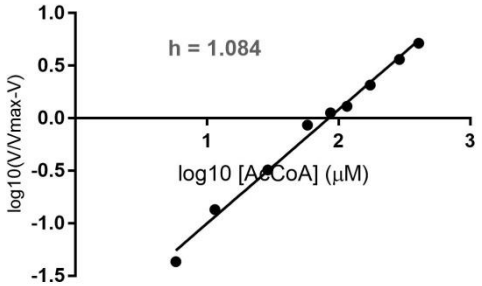


Figure 4.10: Hill plot of *Nme*IPMS K395Term for AcCoA.

4.1.5.5 Michaelis-Menten kinetics in the presence of 30% glycerol

Performing kinetics in the presence of a viscogen such as glycerol can provide information about conformational change and the dynamics of a protein. An example of this type of experiment was provided by Sekhar et al.[155] who used glycerol to probe the viscosity-dependent kinetics of a conformational change in a four helix bundle FF domain. Kinetic analysis was performed in the presence of 30% glycerol to assess the impact of a viscogen on the kinetics and allosteric regulation of *Nme*IPMS, and the kinetics of truncated form of *Nme*IPMS (Table 4.3).

Table 4.3: Kinetic parameters for the wild type *Nme*IPMS and the truncated *Nme*IPMS in the presence and absence of 30% glycerol in the kinetics buffer.*The K_m for KIV was not determined for the truncated *Nme*IPMS. All kinetic analyses were performed at 25°C.

Enzyme	K_m (KIV, μM)	K_m (AcCoA, μM)	k_{cat} (s^{-1})
*Nme*IPMS	36 ± 3	35 ± 3	7.2 ± 0.1
*Nme*IPMS 30% glycerol	57 ± 5	8.9 ± 0.6	1.95 ± 0.02
*Nme*IPMS K395Term	30 ± 3	80 ± 7	4.1 ± 0.1
*Nme*IPMS K395Term 30% glycerol	Nd*	22 ± 2	1.2 ± 0.1

There was a significant 1.6-fold increase in the K_m value for KIV in the wild type protein when 30% glycerol was present in the buffer, but a 3.8-fold decrease in the K_m for AcCoA. The k_{cat} has also decreased from 7 to 1.95 s^{-1}. This implies that there has been a major change in how the protein is operating in the viscous buffer compared to the buffer with aqueous viscosity. The same experiment was conducted with the truncated form of the enzyme, and the K_m for AcCoA also decreased 3.4-fold. The k_{cat} had also decreased compared to the aqueous kinetics, but only 2-fold compared to 3.5-fold for the wild type protein.

Although firm conclusions about the impact of viscosity on the kinetic cycle of *Nme*IPMS cannot be made from these preliminary experiments, as factors such as the size and charge of the viscogen must also be assessed, this data does suggest that viscosity does have a major impact on how the protein catalyses the reaction. To further clarify the impact of viscosity on the kinetic and allosteric behaviour of *Nme*IPMS and other proteins, the same experiments may be performed in the presence of a different viscogen or crowding agent such as bovine serum albumin (BSA). Further experiments, such as performing SAXS in the presence of 30% glycerol, may provide further information about the dynamics of this protein.

Table 4.4: Inhibition data for the inhibition by L-leucine of the wild type *Nme*IPMS in the presence and absence of 30% glycerol.

	IC_{50} (L-leu, μM)	Activity remaining (%)
*Nme*IPMS	55 ± 5	14
*Nme*IPMS 30% glycerol	29 ± 2	19

As the k_{cat} has decreased, this suggests that the product binding or release steps could have been adversely affected by increased viscosity, although it is noted that the rate-determining step under these conditions has not been determined. Most interesting, however, is the considerable impact of viscosity on the K_m for AcCoA. One plausible explanation behind the substantial decrease in K_m for AcCoA for both the wild type and the truncated protein is that the presence of the viscogen limits the conformations that the protein can access, especially the subdomains that are crucial for the binding of AcCoA and are highly mobile, meaning that it is more likely to be in a conformation that can interact with AcCoA.

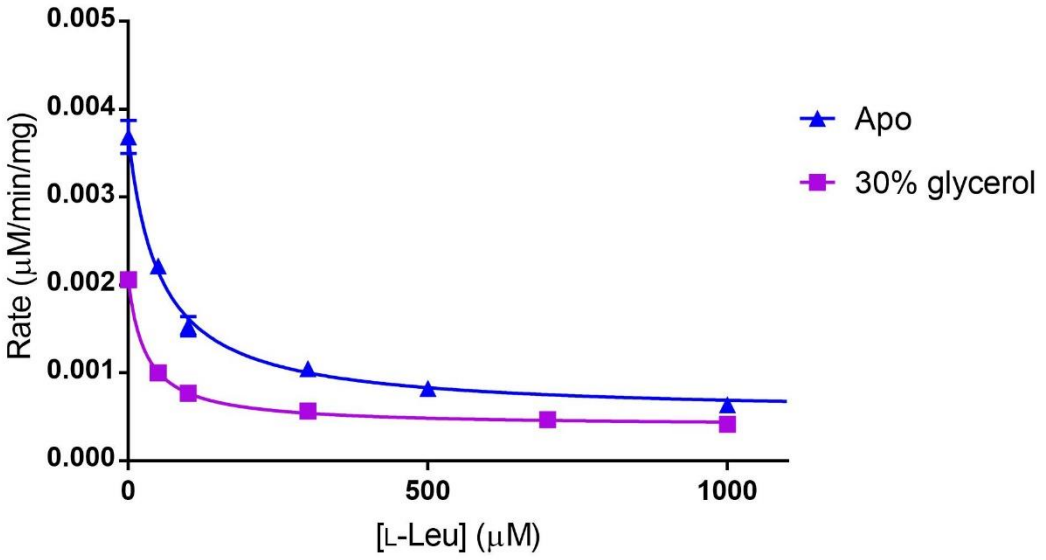


Figure 4.11: The change in the IC_{50} for L-leucine by *Nme*IPMS in the presence and absence of 30% glycerol. The concentration of both substrates was held at 250 μM while the concentration of L-leucine was increased. The response to L-leucine of the wild-type protein in the standard buffer is shown in blue and the response to L-leucine of the wild-type protein in the same buffer containing 30% glycerol is shown in purple.

The inhibition of the wild-type protein by L-leucine was also assessed in the presence of 30% glycerol. There was a small impact on L-leucine inhibition in that the IC_{50} decreased compared to the aqueous buffer, and the residual activity increased (Figure 4.11, Table 4.4). The decrease in IC_{50} suggests that the protein is allosterically inhibited at a lower inhibitor concentration in the presence of 30% glycerol. Inhibition by L-leucine in *Nme*IPMS is reported to be mixed non-

competitive inhibition.[64] The change in IC_{50} could plausibly be due to the restriction of the conformations that the protein can access, meaning that the protein is already limited in the conformational flexibility and thus in a more 'inhibited-like' form even in the absence of inhibitor. This correlates with the decrease in k_{cat} and the decrease in K_m for AcCoA in the presence of the viscogen.

Using viscogens to alter the dynamics of *Nme*IPMS presents a plausible way to decrease the conformations available to the wild type protein and thus explore how this affects catalysis and allosteric regulation. Once the effect of other viscogens, and potentially macromolecular crowding agents such as Ficoll-70, has been explored, other techniques such as ITC could be used to explore the thermodynamics of inhibition in particular. Olsen et al.[156] used BSA, a proteinaceous viscogen, as a macromolecular crowding agent in ITC experiments to assess the impact of a complex environment on the kinetics of hexokinase, which suggests a potential way to explore the thermodynamics of the *Nme*IPMS kinetic cycle.

4.1.6 Conformational dynamics in solution

Small-angle X-ray scattering (SAXS) data was obtained for the apo truncated protein and the truncated protein in the presence of KIV (Figure 4.12). This was obtained to observe the structure of the truncated protein in solution, and any potential conformational change in the presence of the first substrate to bind, KIV.

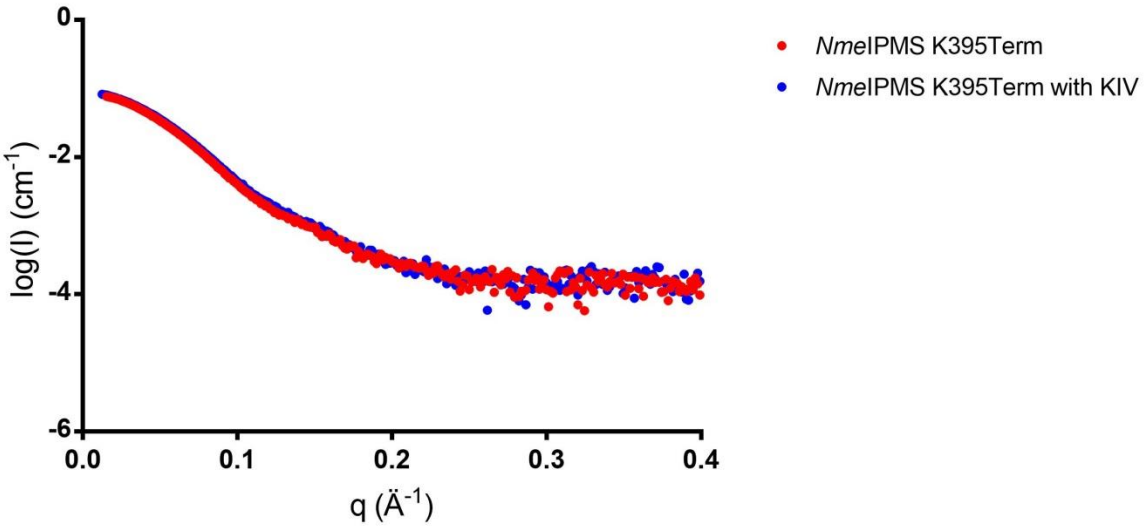


Figure 4.12: Small-angle X-ray scattering data for *Nme*IPMS K395Term in the absence (red) and presence (blue) of the substrate, KIV.The two data sets are shown together, showing the absence of a major conformational change in response to KIV binding.

Table 4.5: SAXS parameters of apo *Nme*IPMS K395Term and KIV-bound *Nme*IPMS K395Term

	Truncated *Nme*IPMS (Apo)	Truncated *Nme*IPMS (200 μM KIV)
I(0) (cm^{-1}) from Guinier plot	0.085 ± 0.00012	0.087 ± 0.00013
I(0) (cm^{-1}) from pairwise distribution function	0.09	0.088
R_g (Å) from Guinier plot	33.8 ± 0.2	33.5 ± 0.3
R_g (Å) from pairwise distribution function	35.5	34.7
I (Å)	140	140
Porod volume estimate ($Å^3$)	149400	138058
Monomeric mass (Da)	46506	46506
Oligomeric structure	Dimer	Dimer

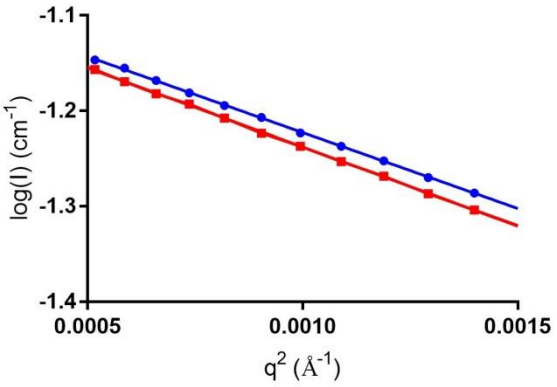


Figure 4.13: Guinier distributions of the SAXS data for apo *Nme*IPMS K395Term (red) and KIV-bound *Nme*IPMS K395Term (blue).

Information obtained from SAXS is dependent on the sample being monodisperse, thus validation of the results obtained is of critical importance. The Guinier distribution was used to assess whether the samples showed significant aggregation. Neither the apo sample (Figure 4.13) or the sample containing KIV (Figure 4.13) showed substantial non-linearity at low q ranges, suggesting that the samples were not aggregating. The radius of gyration (R_g) was also determined from this plot for both samples (Table 4.5).

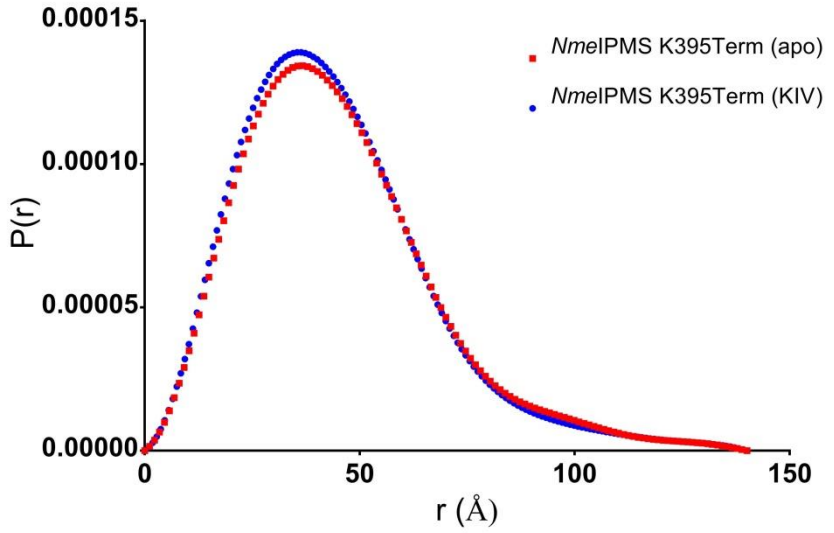


Figure 4.14: Pairwise distributions of apo *Nme*IPMS K395Term (red), and KIV-bound *Nme*IPMS K395Term (blue).

A pairwise distribution function, P(r), was generated for both the apo (Figure 4.14) and KIV bound (Figure 4.14) samples. From these, the R_g and the I(0) could be calculated (Table 4.5). The R_g from

this analysis was compared to the R_g obtained from the Guinier plot. The D_{max}, the maximum dimension of the particle, was also obtained from the P(r) distribution, as was the Porod estimate of excluded volume that gives an indication of the molecular mass (Table 4.5). The R_g and I(0) calculated from the two different methods for both samples are similar, suggesting, as with the Guinier plots, that there is not significant aggregation in the samples.

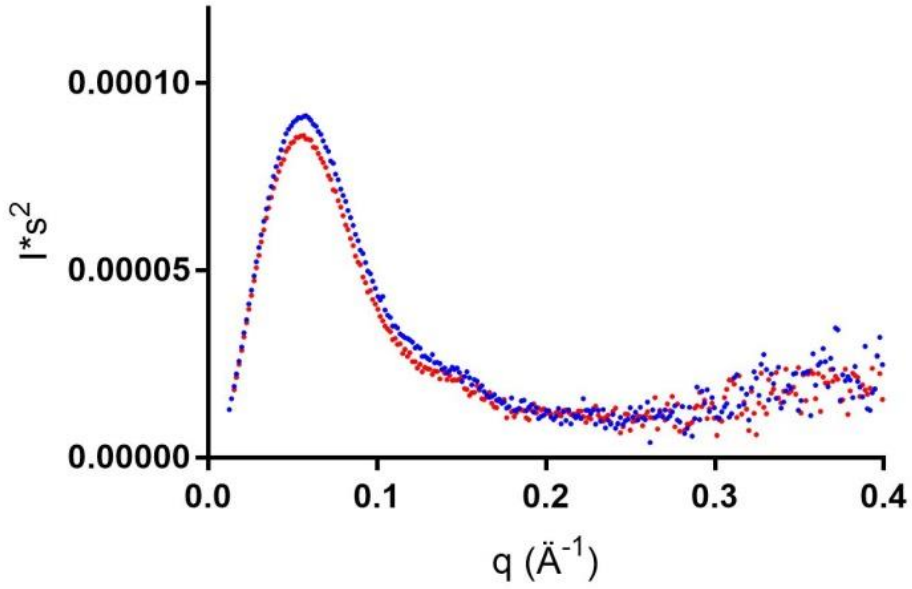


Figure 4.15: Kratky plot of apo *Nme*IPMS K395Term (red) and KIV-bound *Nme*IPMS K395Term (blue)

Kratky plots (Figure 4.15) of both samples also suggest that there is substantial flexibility in both of the samples although there is a significant well folded component. The shape of the Kratky plots suggests that the truncated protein is more compact in the presence of KIV. This information corresponds well with the premise that, while unrestrained by the absence of a regulatory domain, the subdomains can form multiple conformations in solution. The presence of KIV appears to decrease the flexibility of the protein, suggesting that there is a change in dynamics in the presence of KIV.

4.1.6.1 Model fitting

Crysol was used to fit homology models of the truncation that were made by Dr. Wanting Jiao (personal communication, November 2016). The models were constructed using molecular dynamics, having removed the regulatory domain. The starting conformation ("start" model) for the MD simulation included the position of the subdomains in the full length *Nme*IPMS homology model which was allowed to proceed until energy was minimised. The "start" model (Figure 4.16, top left), and "end" model (Figure 4.16, top right) were used as input in the Crysol programme to fit the theoretical scattering of these models to the scattering produced by the *Nme*IPMS truncation (Figure 4.16, middle and bottom). The theoretical scattering produced by either model does not fit the scattering produced by the *Nme*IPMS K395Term truncation as indicated by the high chi^2 values.

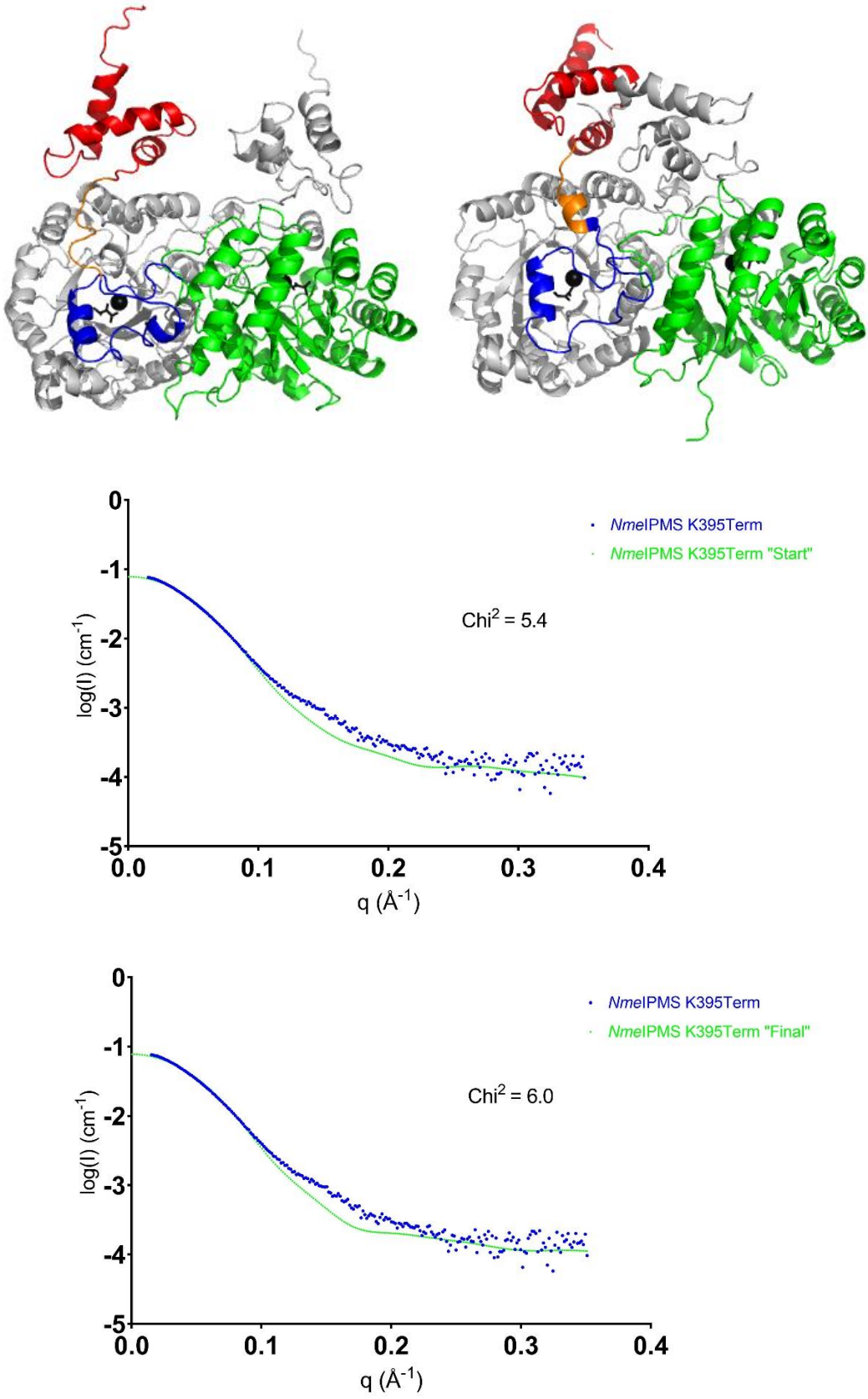


Figure 4.16: Top: Models of the *Nme*IPMS K395Term truncation. *Nme*IPMS K395Term "Start" is shown on the left, and *Nme*IPMS K395Term "Final" is shown on the right. Chain A is shown in grey, the catalytic domain of chain B is shown in green, subdomain I in blue, the linker in orange, and subdomain II in red. Middle: The "start" model theoretical scattering (green) fitted

to the scattering produced by *Nme*IPMS K395Term (blue). Bottom: The "final" model theoretical scattering (green) fitted to the scattering produced by *Nme*IPMS K395Term (blue).

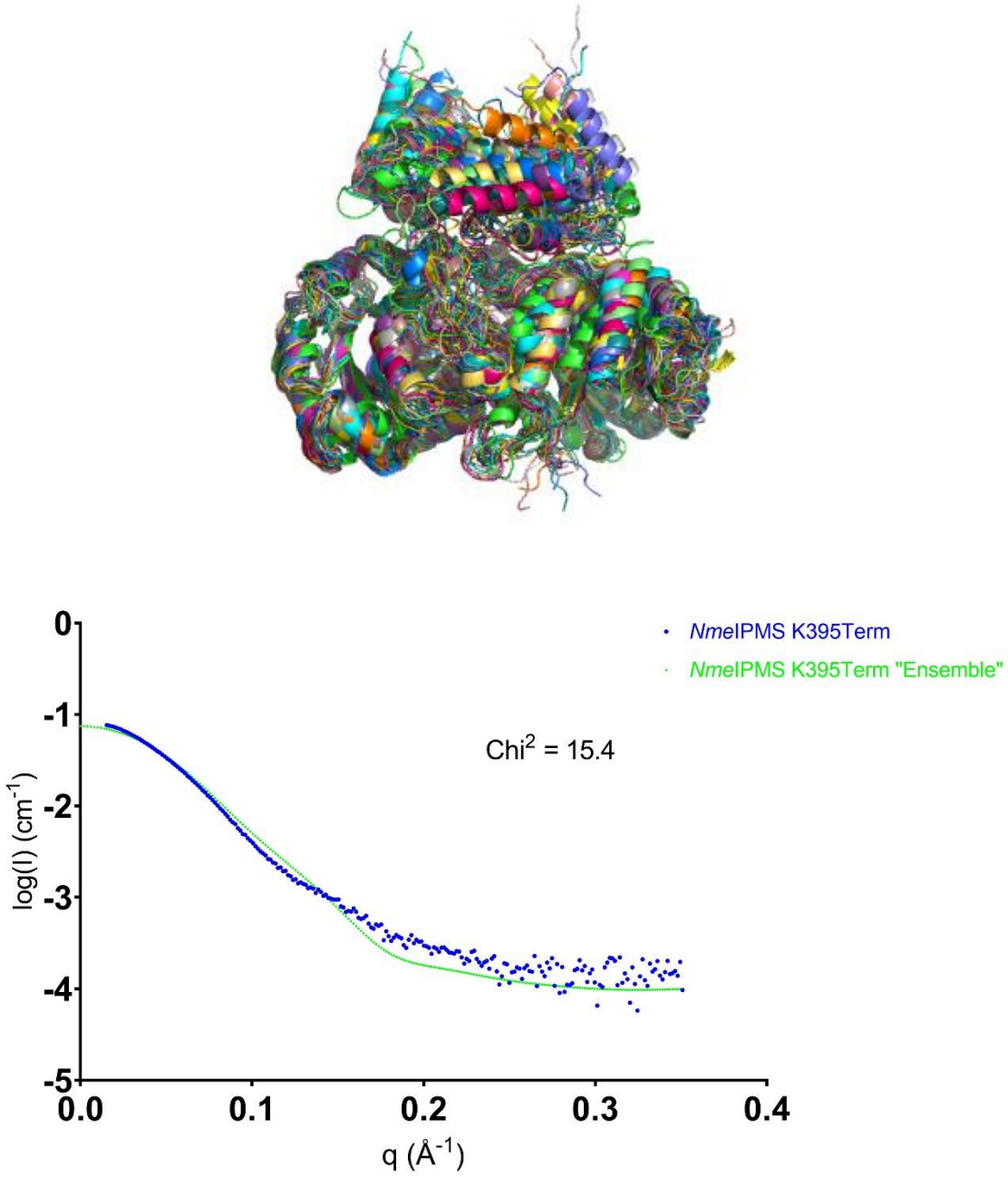


Figure 4.17: The scattering of an ensemble of models of *Nme*IPMS K395Term compared to the SAXS scattering of *Nme*IPMS K395Term.Top: An ensemble of models of *Nme*IPMS K395Term generated from molecular dynamics simulations. Bottom: The theoretical scattering of the above ensemble (green) fitted to the scattering produced by *Nme*IPMS K395Term (blue).

Crysol was used to fit the theoretical scattering profiles of ensembles of models generated by molecular dynamics to the scattering profile of the samples (Figure 4.17).[157] Neither the single conformations nor the ensemble fit the scattering profile, suggesting that the protein may access additional conformations that are not observed during this specific MD simulation. As the subdomains are un-restrained at one end, there is a considerable likelihood that, particularly in the

absence of ligand, the subdomains can sample conformational space to a large degree. Additionally, structural data for either the full length *Nme*IPMS or active truncated *Nme*IPMS has not been obtained, so the theoretical scattering profiles are generated from MD performed on a homology model that was constructed based on the *Mtu*IPMS full length crystal structure that shows considerable difference from *Nme*IPMS in sequence and evolution. It is plausible that the models used to generate the theoretical scattering do not adequately describe the truncated form of *Nme*IPMS.

4.1.7 SAXS of the wild type *Nme*IPMS

Small-angle X-ray scattering data was obtained for the wild type *Nme*IPMS in its apo form at pH 7.5 (Figure 4.18). Attempts were made to obtain SAXS data in the presence of L-leucine at several concentrations. However, under all conditions tested, there was significant aggregation in the sample. This problem has been noted in the past.[64]

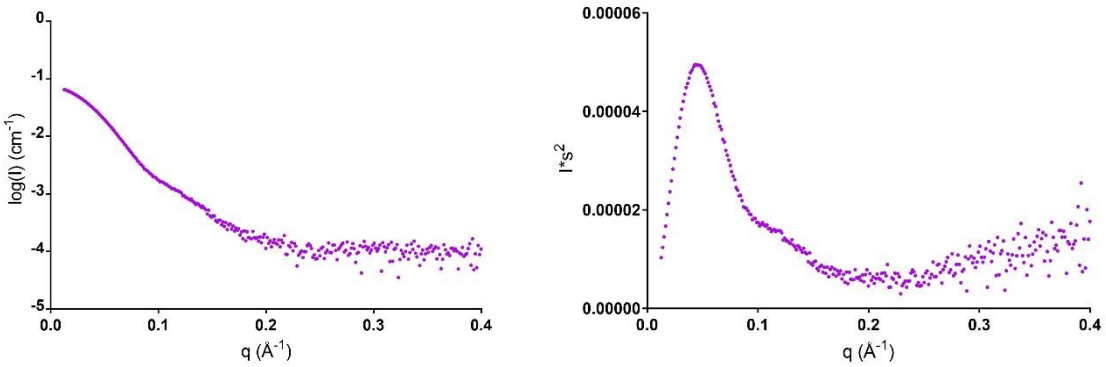


Figure 4.18: SAXS data for apo *Nme*IPMS WT (left) and a Kratky plot of the same data (right)

The Guinier plot of the apo data shows that there is no substantial aggregation (Figure 4.19). The R_g calculated from the Guinier plot and that calculated from the pairwise distribution (Figure 4.19) correspond well, providing further evidence that there is no substantial aggregation in the sample.

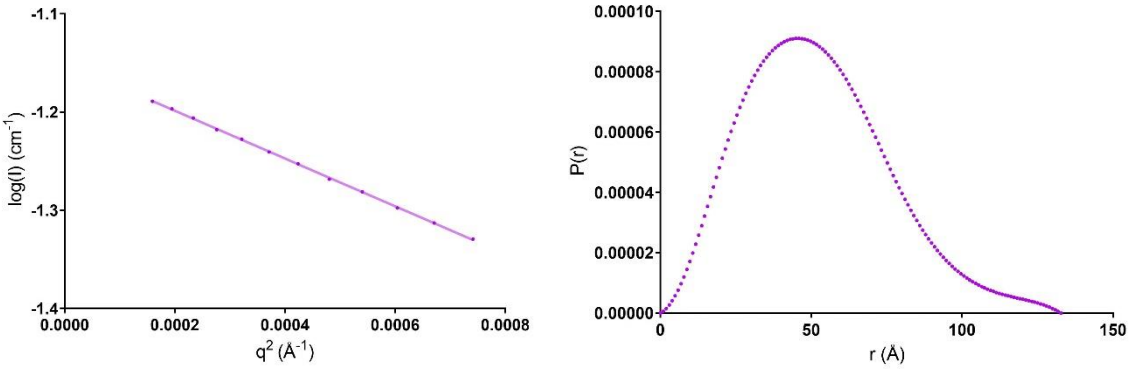


Figure 4.19: Guinier distribution and pairwise distribution of apo *Nme*IPMS WT

The pairwise distribution function suggests that the wild type apo protein is primarily globular in solution. The Kratky plot (Figure 4.18) suggests that there are multiple domains in the protein and substantial flexibility. In the absence of leucine, it has been suggested that the subdomains and

regulatory domain form multiple conformations although these conformations have not been structurally assessed. The information obtained from SAXS also suggests this.

Theoretical scattering of the *Nme*IPMS homology model, and of ensembles generated from molecular dynamics simulations, was also produced and fitted to the data using Crysol. The theoretical scattering of the homology model did not fit the scattering of the protein well, with a chi^2 value of 15.01.

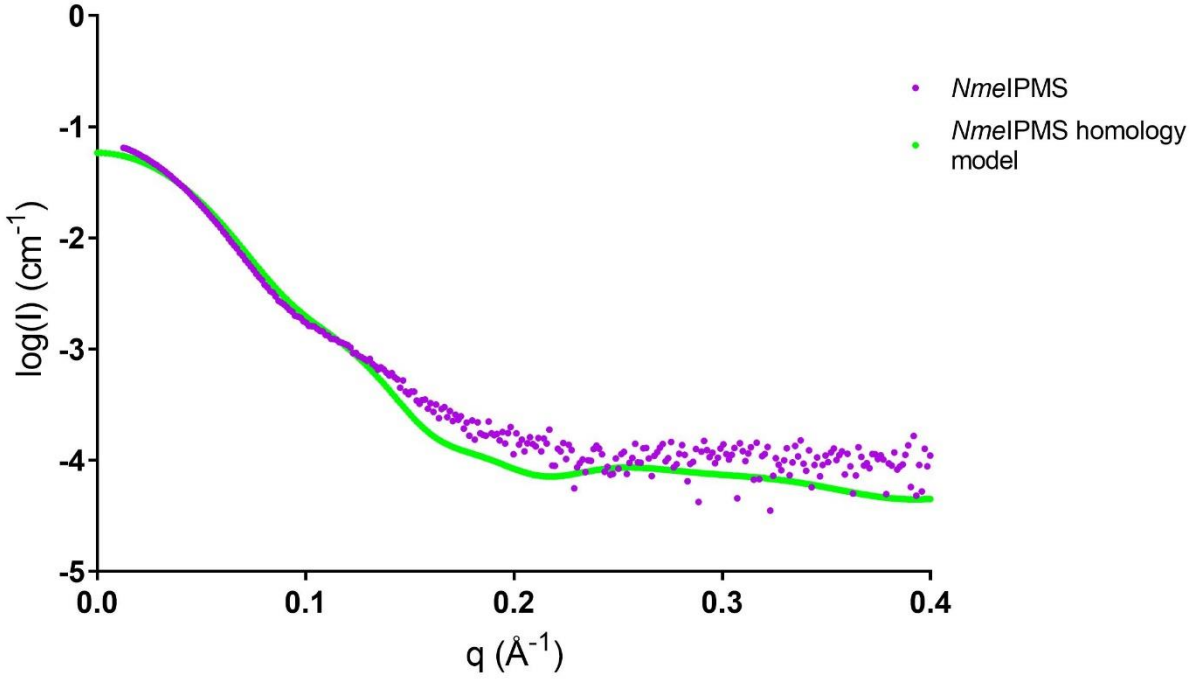


Figure 4.20: The *Nme*IPMS homology model theoretical scattering (light green) fitted to the scattering produced by *Nme*IPMS (purple). The chi^2 value of the fit was 15.01.

Table 4.6: SAXS parameters of *Nme*IPMS wild type

	*Nme*IPMS (Apo)
I(0) (cm^{-1}) from Guinier plot	0.071 ± 0.00016
I(0) (cm^{-1}) from pairwise distribution function	0.07
R_g (Ä) from Guinier plot	40.96 ± 0.64
R_g (Ä) from pairwise distribution function	40.23
D_{max} (Ä)	133
Porod volume estimate ($Å^3$)	237545
Monomeric mass (Da)	59208
Oligomeric structure	Dimer

4.1.8 Crystallography

Crystallisation of *Nme*IPMS K395Term was attempted using both sitting-drop vapour diffusion and hanging-drop vapour diffusion under a variety of conditions. Sitting drop vapour diffusion conditions included commercial screens: JCSG+ (Molecular Dimensions), Clear Strategy 1 (Molecular Dimensions), Clear Strategy 2 (Molecular Dimensions), and PACT (Molecular Dimensions). Protein concentrations of 5 mg/mL to 25 mg/mL were used, and screens were constructed with and without the ligand, KIV. A Mosquito® crystallisation robot (TTP LabTech) was used to put down the sitting drop vapour diffusion screens. A screen of conditions based around the crystallisation condition that produced crystals of *Nme*IPMS Glu365Term[158] was also constructed. This screen varied protein concentration (10 mg/mL and 15 mg/mL), magnesium acetate concentration (0.1 M to 0.25 M), and PEG3350 (10% to 20%).

Very little precipitation was seen at protein concentrations under 20 mg/mL, suggesting the protein is very soluble under these conditions. However, no lead crystals were seen under any conditions. The full-length protein has not been crystallised, although attempts have been made in the presence and absence of ligands including L-leucine. A crystal structure of the E365Term truncation of *Nme*IPMS has been solved[158] but as the full length and active truncated proteins are very flexible and very dynamic, they may prove difficult to crystallise.

4.1.9 Alanine mutants of *Nme*IPMS wild type and *Nme*IPMSK395Term

Several residues in *Nme*IPMS, namely Tyr313, Lys332, and Arg371 have been identified as potentially important for catalytic activity in previous research.[131, 132] Davies determined that the *Nme*IPMS Tyr313Phe and Lys332Ala mutants showed catalytic activity but further characterisation performed by Plowman-Holmes[131] suggested that these mutants did not show catalytic activity although these assays were performed only with a limited amount of protein. When a large amount of protein was used in the kinetic assay, the mutant proteins did display some catalytic function. To explore the roles of these residues in providing catalytic function, kinetic parameters and whether the mutant proteins were inhibited by L-leucine were determined.

Arg371, located in subdomain II, was identified as a residue that may be involved in the recruitment of AcCoA to the active site (Dr. Wanting Jiao, personal communication, May 2014). To investigate the role of this residue in the catalytic cycle, an alanine mutant, Arg371Ala, was made in both the full-length wild-type protein and the truncated K395Term protein.

Table 4.7: Kinetic and inhibition parameters of *Nme*IPMS WT and *Nme*IPMS mutants. In this table, N/A stands for not applicable, and nd stands for not determined. The IC_{50n} for *Nme*IPMS K332A and *Nme*IPMS Y313F was not determined as the k_{cat} of the protein was too low to obtain this data.

	Apparent K_m (KIV) (μM)	Apparent K_m (AcCoA) (μM)	k_{cat} (s^{-1})	Leu Sensitive?	IC_{50} for Leu (μM)	Residual activity (%) at full inhibition
*Nme*IPMS WT	36 ± 3	35 ± 3	7.2 ± 0.1	Yes	53 ± 5	17
*Nme*IPMS R371A K395Term	70 ± 5	220 ± 20	1.83 ± 0.07	No	N/A	N/A
*Nme*IPMS K332A	Nd	900 ± 100	0.15 ± 0.01	Yes	nd	nd
*Nme*IPMS R371A	Nd	220 ± 20	3.8 ± 0.1	Yes	45 ± 5	12
*Nme*IPMS Y313F	490 ± 50	35 ± 3	0.0078 ± 0.0001 and 0.110 ± 0.005	Yes	nd	nd

4.1.9.1 *Nme*IPMS Tyr313Phe

IPMS and related enzymes contain a conserved tyrosine in subdomain I that inserts into the active site of the opposite chain. In *Mtu*IPMS, this tyrosine (Tyr410) stacks with His379, also in subdomain I, potentially to position this latter residue as the catalytic base (Figure 4.21).[67] Glu218 from the other chain is also in a position to act as the catalytic base. Alanine mutants of Glu218 and His379, and a phenylalanine mutant of Tyr410, have been made in *Mtu*IPMS.[63] *Mtu*IPMS Glu218Ala and *Mtu*IPMS His379Ala both showed aberrant kinetics, and both displayed substrate activation towards KIV, whereas only Glu218Ala showed substrate activation towards AcCoA as well.[63] Although the Tyr410Phe *Mtu*IPMS mutant displayed a large decrease in k_{cat}, there was also a corresponding decrease in the K_m for both substrates, suggesting that this mutation allows the protein to bind substrates more tightly. Unlike the His379Ala or Glu218Ala mutants, *Mtu*IPMS Tyr410Ala was insensitive to inhibition by L-leucine.[63]

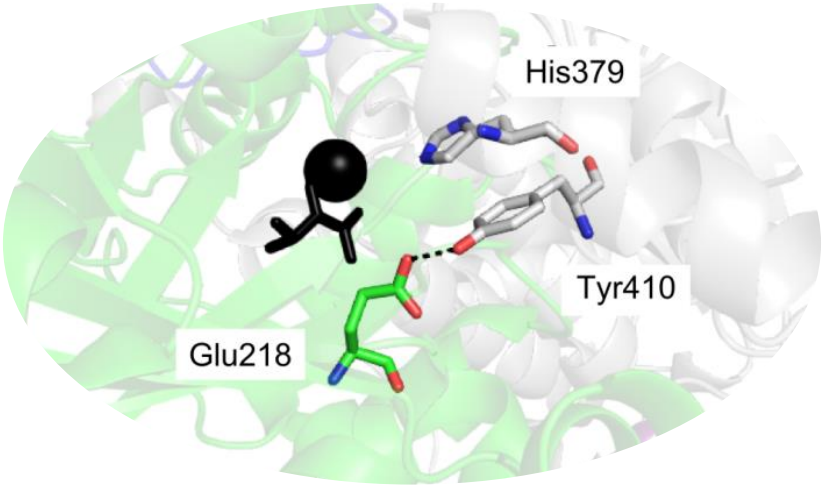


Figure 4.21: The location of Tyr410, His379, and Glu218 in the *Mtu*IPMS structure (PDB: 1SR9). KIV is shown as black sticks and the metal ion as a black sphere. The stacking between His379 and Tyr410, and the interaction between Glu218 and Tyr410, that orients Tyr410 correctly in the active site can be seen.

To explore the role of the conserved tyrosine residue, in a key location in subdomain I, in the catalytic cycle and allosteric regulation of *Nme*IPMS, a phenylalanine mutant of Tyr313, the corresponding residue in *Nme*IPMS, was made.[132] The initial characterisation of this mutant was performed by Plowman-Holmes[131].

Michaelis-Menten kinetic data were obtained for this mutant (Figure 4.22, Table 4.7). *Nme*IPMS Tyr313Phe, similarly to *Mtu*IPMS Tyr410Phe, showed a substantial decrease in k_{cat} compared to

the wild type protein. The k_{cat} was determined both from the experiment where the concentration of AcCoA was altered, where it was 0.0078 s^{-1} and when KIV was altered, where it was 0.11 s^{-1}. As the apparent K_m for KIV was so high, the concentration of KIV could not be held at a saturating level when the apparent K_m for AcCoA was determined, and this may account for the difference in rates seen in Figure 4.22 and subsequently the difference in k_{cat} between the two experiments. The extremely high K_m for KIV also meant that the V_{max} in the experiment where KIV concentration was varied could not be reliably determined. This problem also suggests that the k_{cat} determined from the variation in AcCoA concentration is likely to be artificially low.

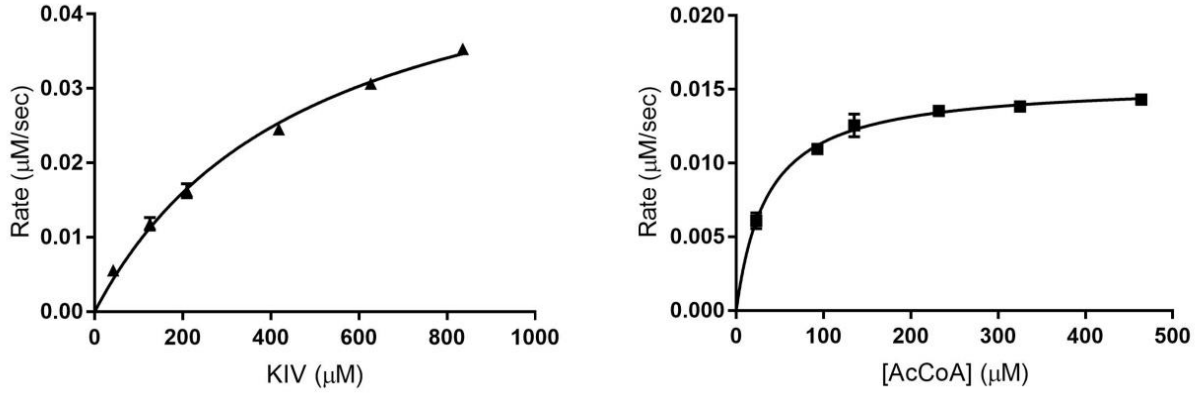


Figure 4.22: Michaelis-Menten plots for *Nme*IPMS Tyr313Phe for the substrates KIV (left) and AcCoA (right).When the kinetic parameters of *Nme*IPMS Tyr313Phe were determined, the concentration of [AcCoA] was held at 230 μM while the concentration of KIV was altered, and when the concentration of AcCoA was altered, the concentration of KIV was held at 420 μM.

Unlike *Mtu*IPMS Tyr410Phe, the K_m for AcCoA was similar to that of wild type *Nme*IPMS, suggesting that this mutation has not adversely affected interaction with AcCoA. Additionally, the K_m for KIV for *Nme*IPMS Tyr313Phe had increased 14-fold compared to the wild type protein, suggesting that this mutation drastically decreases the ability of the protein to interact with KIV. As the Tyr410Phe mutation in *Mtu*IPMS had the opposite effect on the K_m for KIV, this suggests that the two proteins may have different interactions with the substrates in the active site.

L-Leucine sensitivity of the mutant was also assessed, and unlike *Mtu*IPMS Tyr410Phe, the *Nme*IPMS Tyr313Phe was sensitive to L-leucine (Table 4.7, Figure 4.23). This suggests, as discussed in Chapter 2, that there are different allosteric pathways conferring allosteric regulation by L-leucine in different members of the IPMS family. The mutant enzyme shows a considerably reduced response to L-leucine compared to the wild-type protein (Figure 4.23). However, this experiment was carried out under saturating conditions for the wild-type protein that are not

saturating, with respect to KIV, for the mutant enzyme. Under saturating conditions for KIV, the response to L-leucine may be different in the Tyr313Phe mutant.

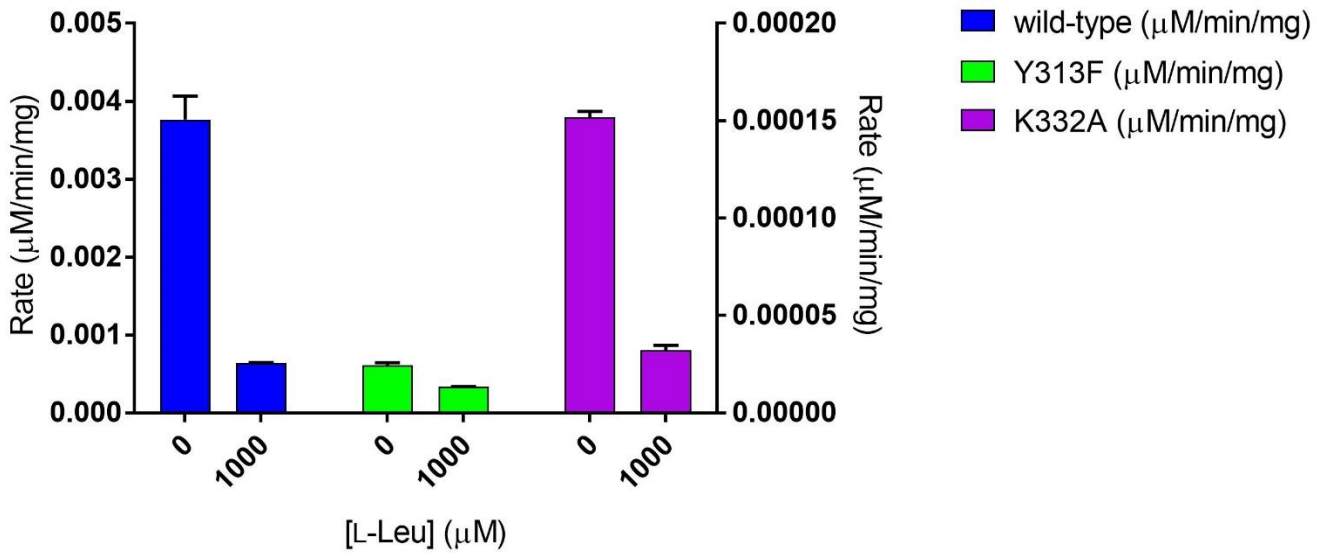


Figure 4.23: The response of *Nme*IPMS wild-type, *Nme*IPMS Tyr313Phe (Y313F), and *Nme*IPMS Lys332Ala (K332A) to L-leucine. Due the low k_{cat} of both Y313F and K332A, a complete IC_{50} plot for the mutant could not be completed. The concentration of AcCoA was held at 230 μM and the concentration of KIV was held at 210 μM in all analyses. The specific activity of *Nme*IPMS wild-type (blue) is plotted on the left axis and the specific activity of Y313F and K332A are plotted on the right axis.

4.1.9.2 *Nme*IPMS Lys332Ala

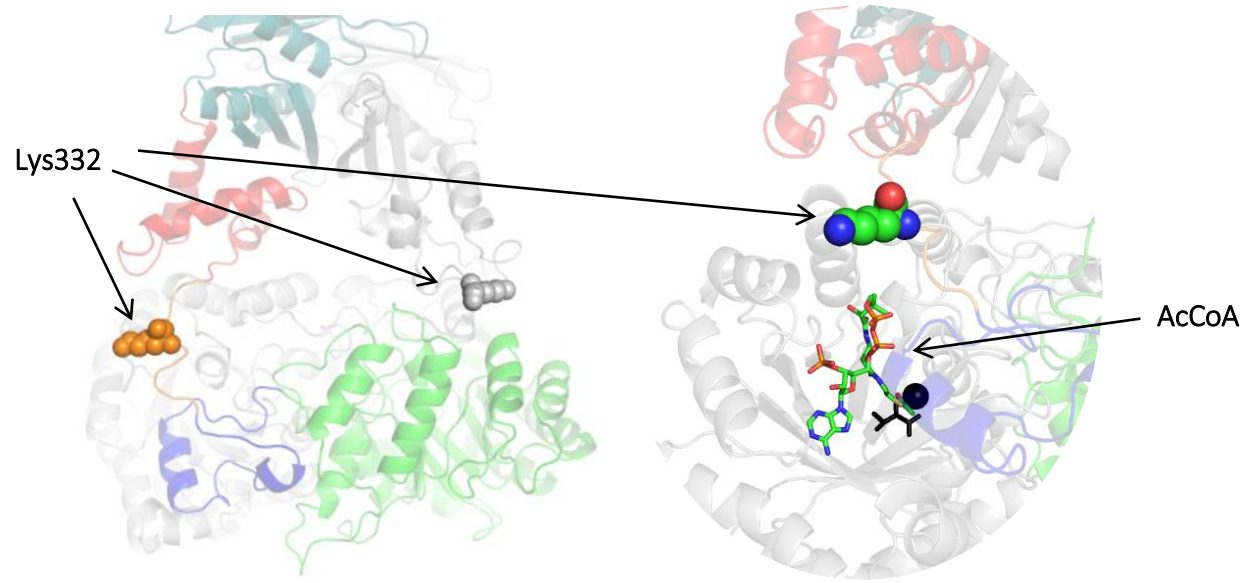


Figure 4.24: The location of residue Lys332 in the *Nme*IPMS homology model. Chain A is shown in grey, while chain B is shown in green (catalytic domain), blue (subdomain I), orange (linker), red (subdomain II), and teal (regulatory domain). The right-hand image shows Lys332 (spheres, coloured by element) in the linker and AcCoA obtained from the 3BLI structure of *Lin*CMS that was structurally aligned with the *Nme*IPMS homology model.

Molecular dynamics combined with docking of AcCoA performed by Dr. Wanting Jiao (personal communication, October 2015) identified several residues in the subdomains that may have a role in AcCoA recruitment. This set of residues includes residue Lys332 in *Nme*IPMS, which is located on the loop between subdomains I and II. This region has conserved positive charge, and in the RDP alignment, this residue is almost absolutely conserved. This suggests that this residue has a critical role in catalysis.

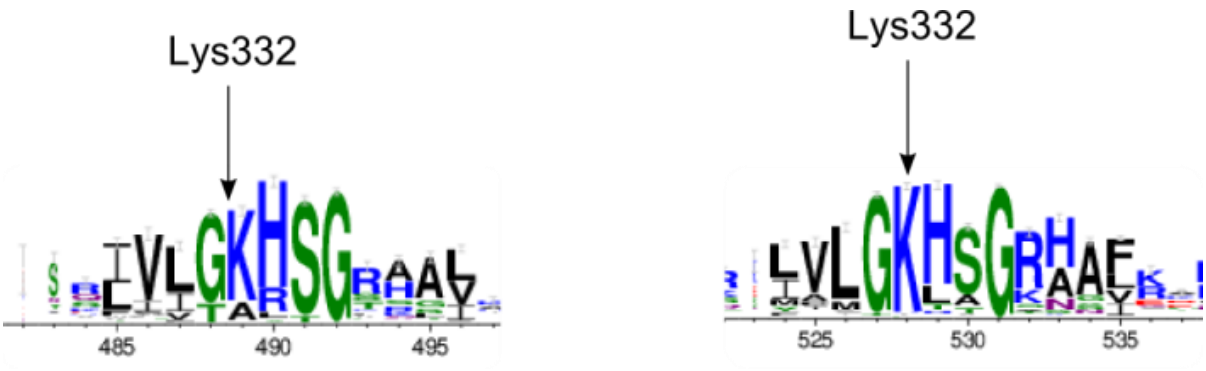


Figure 4.25: Logo diagram of the regulatory-domain absent (RDA) (left) and regulatory domain present (RDP) (right) alignments from Chapter 3 showing the conservation of residue Lys332 (*Nme*IPMS numbering) in both alignments. The x-axis numbering refers to the residue number in the alignment.

An alanine mutant, Lys332Ala, was made. The mutant was constructed by Davies[132], and initial characterisation was performed by Plowman-Holmes[131]. This mutant displayed a large increase in the apparent K_m for AcCoA; the wild-type protein has an apparent K_m for AcCoA of 35 ± 3 μM while the Lys332Ala mutant has an apparent K_m for AcCoA of 900 ± 100 μM. This enormous change in apparent K_m demonstrates that this residue is critical for the interaction with AcCoA even though, in the homology model, it does not form part of the active site (Table 4.7, Figure 4.26). The K_m for KIV could not be determined as the K_m for AcCoA was so high that the concentration of AcCoA could not be held at or near saturation.

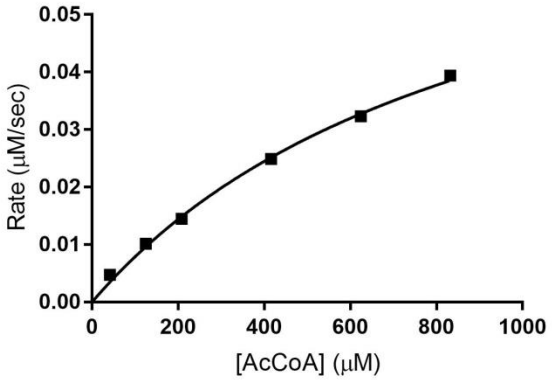


Figure 4.26: Plot of the kinetic data of *Nme*IPMS Lys332Ala where the concentration of AcCoA is varied. The substrate concentration of KIV was held at 250 μM

The importance of positive charge in this location to catalytic activity may provide some understanding as to why subdomain II is essential for catalysis. The increase in K_m for AcCoA in the Lys332Ala mutant suggests that this residue is involved in the recruitment of AcCoA. Lys332 is located in the flexible loop between subdomains I and II, suggesting that the restriction on the conformations available to this region determined by the three-helix bundle of subdomain II allows the loop to be available to recruit AcCoA. In the absence of subdomain II, or when it has been partially truncated, this restriction has been abolished and the loop is not restrained sufficiently to allow for AcCoA recruitment to the active site. A k_{cat} of 0.15 ± 0.01 s^{-1} was estimated from the kinetic data obtained in Figure 4.26, although due to the high K_m of this enzyme, this is likely to be considerably higher than the actual turnover number. This has significantly decreased compared to the kcat of the wild-type protein which was determined to be around 7 s^{-1}.

Leucine sensitivity in this alanine mutant was also investigated (Figure 4.23, Table 4.7). This mutant, like Tyr313Phe, also shows sensitivity to leucine. As with the Tyr313Phe mutant, the IC_{50} of the protein could not be obtained due to the extremely high K_m for at least one of the substrates. Additionally, the very low k_{cat} for both the Tyr313Phe and Lys332Ala mutant made determining changes in rate especially in the presence of an inhibitor challenging as a large amount of protein was required. A true comparison between the residual activity of the mutant compared to wild type in the presence of inhibitor could also not be made, as the substrate concentrations used in the presence of the inhibitor were saturating for the wild-type protein but not for the mutant. However, the sensitivity of both these mutants to L-leucine even though they display extremely aberrant kinetic activity suggests that the allosteric network may utilise interactions beyond those that are important for catalytic function.

4.1.9.3 *Nme*IPMS Arg371Ala and *Nme*IPMS K395Term Arg371Ala

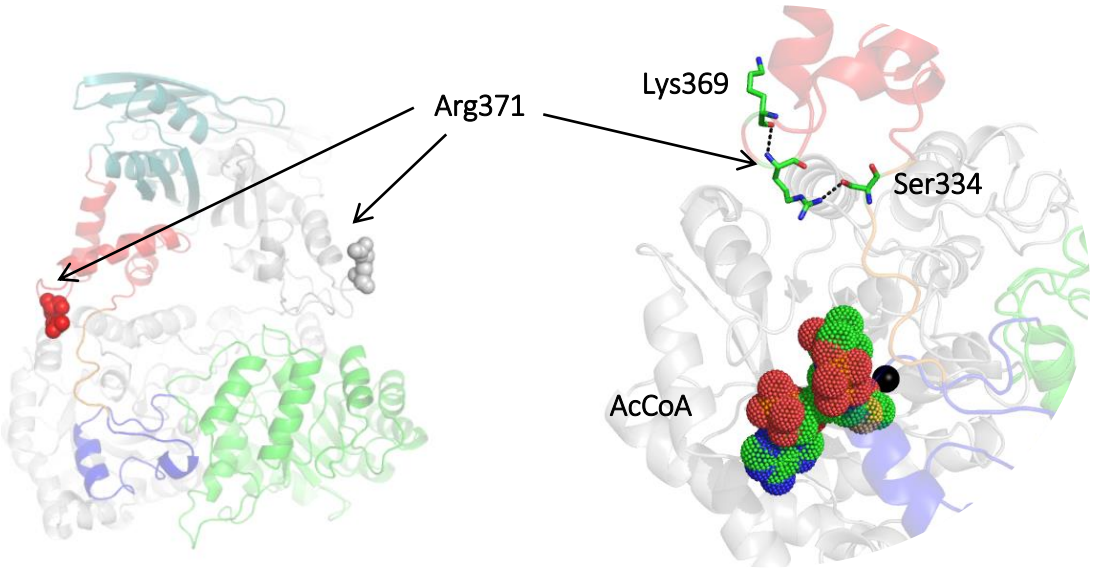


Figure 4.27: The location of Arg371 (spheres) in the *Nme*IPMS homology model. Chain A is shown in grey. Chain B is shown in green (catalytic domain), blue (subdomain I), orange (linker), red (subdomain II), and teal (regulatory domain). The right-hand image shows chain B of the *Nme*IPMS homology model with AcCoA (dots) from the *Lin*CMS structure (PDB: 3BLI) superimposed onto it. The backbone of Arg371 in Chain B forms an interaction with the backbone of Lys369 and the sidechain of Arg371 also forms an interaction with the sidechain of Ser334.

Arg371, like Lys332, was also identified as a residue that forms interactions with AcCoA in MD docking simulations performed by Dr. Wanting Jiao (personal communication, October 2015). Arg371 is found in subdomain II, and, like Lys332, is located in a loop, although Arg371 is located in a loop between helices of subdomain II, while Lys332 is located in a flexible loop between the subdomains. An alanine mutant of this residue was made in the full-length protein, and in the truncated protein, to determine the role of this residue in the recruitment of AcCoA in the presence and absence of a regulatory domain.

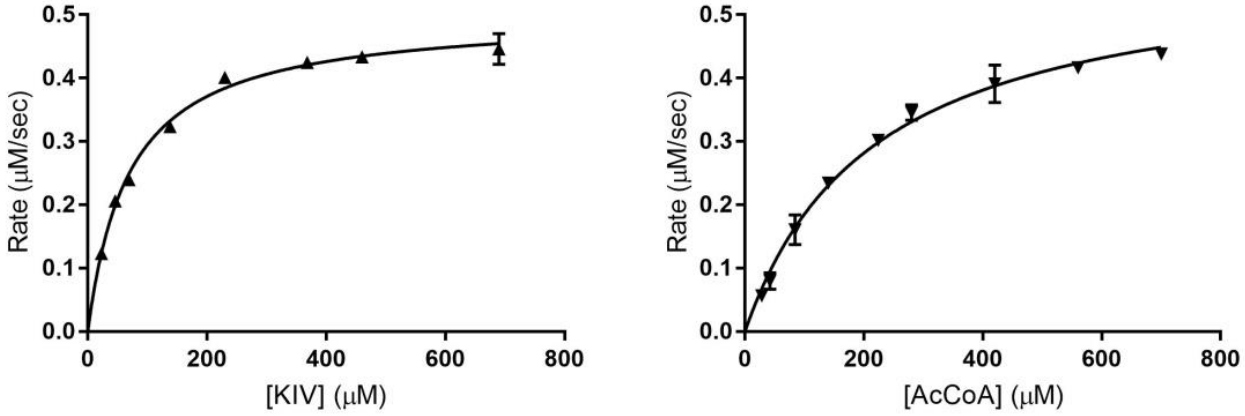


Figure 4.28: Plots of kinetic data for *Nme*IPMS Arg371Ala Lys395Term, showing the change in initial rate when substrate concentration is increased.The data for KIV is shown on the left, and the data for AcCoA is shown on the right. The concentration of AcCoA was held at 300 μM while the concentration of KIV was varied, and the concentration of KIV was held at 230 μM while the concentration of AcCoA was varied.

The full length Arg371Ala mutant showed a substantial increase in the K_m for AcCoA, from 35 μM to 220 μM, although less than that seen in the Lys332Ala mutant which had a K_m for AcCoA of ~900 μM (Table 4.7, Figure 4.28). This increase suggests that this residue is important for the recruitment of AcCoA in the full-length protein. As with Lys332Ala and Tyr313Ala, this mutant is sensitive to leucine. The IC_{50} for leucine was determined, and was comparable to the wild type *Nme*IPMS, as the IC_{50} for wild-type *Nme*IPMS for L-leucine is 53 ± 5 μM and the full-length Arg371Ala mutant had an IC_{50} for L-leucine of 45 ± 5 μM. The lack of change to the IC_{50} value in the mutant protein compared to the wild-type protein suggests that this residue, and the interaction(s) it forms, are not involved in transmitting the allosteric signal to the active site.

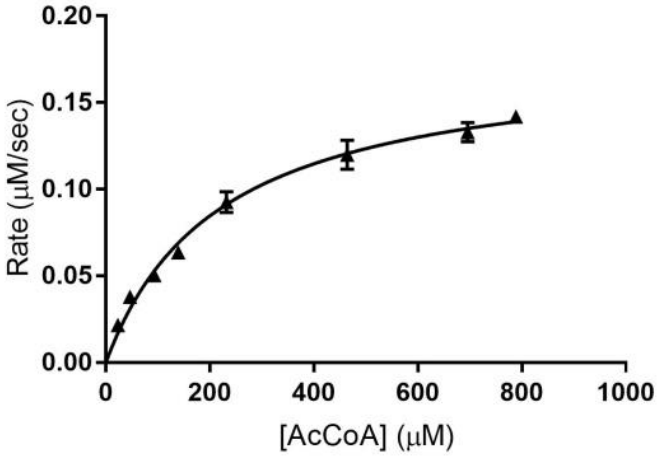


Figure 4.29: Plot of the kinetic data of *Nme*IPMS Arg371Ala in response to change in AcCoA concentration. The concentration of KIV was held at 230 μM as the concentration of AcCoA was varied.

The K_m for AcCoA for the *Nme*IPMS Arg371Ala K395Term mutant was also increased compared to the wild type *Nme*IPMS, and also increased compared to the *Nme*IPMS truncation (Table 4.7, Figure 4.29). Interestingly, the K_m for this mutant is the same as the K_m for the full-length mutant, even though the K_m for AcCoA for *Nme*IPMS K395Term is 2-fold higher than that of the full-length protein, suggesting that this mutation affects the interaction with AcCoA but not the conformations that the subdomains can adopt. The apparent K_m for KIV for this mutant was obtained, and it had increased compared to the wild type protein, although this may have been because the concentration of AcCoA was not saturating. In the future, the K_m for KIV could be more accurately determined by determining the K_m for KIV at different, non-saturating, AcCoA concentrations.

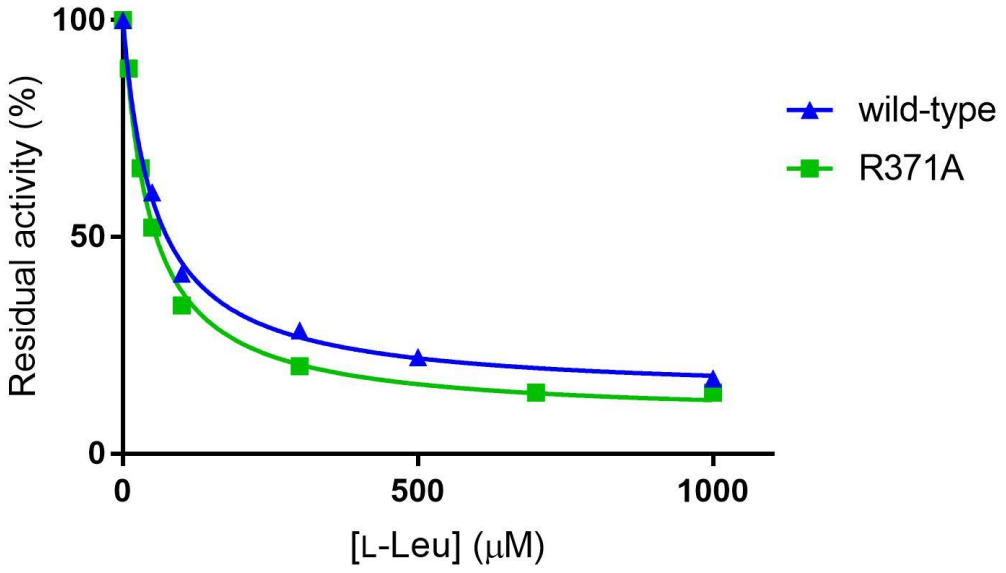


Figure 4.30: Inhibition of *Nme*IPMS Arg371Ala by L-leucine. The concentration of the substrates was held at 300 μM (AcCoA), and 230 μM (KIV). The wild-type *Nme*IPMS IC_{50} data is shown in blue and the Arg371Ala IC50 data is shown in green.

4.2 Discussion

Previous studies have shown that subdomain II is crucial for catalysis in IPMS from several organisms, even though it is located distant to the active site, and it has been have suggested that the role of subdomain II in catalysis is in the recruitment of AcCoA.[2] It has also been suggested that the balance of stability and flexibility in the subdomains is important for catalysis, as sufficient stability is required to restrain the subdomains into conformations that allow for the recruitment

of AcCoA by the subdomains and linker to the active site, while allowing sufficient freedom to explore those conformations.

A truncation in *Nme*IPMS was made to encompass all of subdomain II as a previous truncation, reported to be inactive, truncated the protein midway through subdomain II. This truncation was catalytically active yet insensitive to inhibition by leucine, although the K_m for AcCoA had increased 2-fold compared to the wild type protein. This increase in K_m suggests that removal of the regulatory domain had adversely affected the interaction of the protein with the AcCoA substrate and thus affected catalysis.

SAXS data suggested that the subdomains were highly mobile in solution, and kinetics performed in the presence of a viscogen showed a decreased apparent K_m for AcCoA for both the full length and truncated proteins, suggesting that restraining the conformations that the subdomains can explore is crucial for the recruitment of AcCoA.

Additionally, several alanine mutants of residues in the subdomains were constructed to further explore the role of particular residues in catalysis, and particularly, the recruitment of AcCoA to the active site. The results obtained from these mutations suggest that the dynamics of the subdomains, as well as particular chemistry within those subdomains, are crucially important for the recruitment of AcCoA in particular. The Tyr313Phe mutation in subdomain I suggests that this residue is important for interaction of KIV to the active site, and unlike the comparable mutation in *Mtu*IPMS, mutation of this residue does not abolish sensitivity to leucine.

The mutation of *Nme*IPMS Lys332Ala showed that positive charge in the linker region between subdomains I and II is particularly important for the recruitment of AcCoA, while the Arg371Ala mutant suggested that positive charge in the loop between helices in subdomain II may also be important for recruitment of AcCoA to the active site. A plausible theory based on these results suggests that the role of subdomain II in catalysis is to facilitate the recruitment of AcCoA to the active site, and the ability of the subdomains to move dynamically is crucially important for this function.

The IPMS and IPMS-like proteins with and without a regulatory domain present an interesting conundrum: how can the similarly structured subdomains manage to provide sufficient stability yet flexibility to allow for the linker to recruit AcCoA to the active site in both the presence and absence of a regulatory domain? These results suggest that subdomain II provides sufficient restriction of conformations that the subdomains can form in the absence of a regulatory domain, even in a protein that has evolved to bear the burden of a regulatory domain, to facilitate catalysis.

Interestingly, the K_m for AcCoA decreased substantially in the presence of a viscogen, suggesting that further restricting the subdomains by reducing the flexibility of the protein in solution allows for more efficient interaction with the substrate.

Future work in this area could encompass the results of Chapters 3 and 4, and further explore the residues that demonstrate substantial coevolution in the subdomains in the presence or absence of a regulatory domain and that show differences between the regulatory domain presence and regulatory domain absent populations. One example of this is Phe363 (*Nme*IPMS numbering) that forms part of the hydrophobic core of the three-helix bundle of subdomain II. This residue is typically large and hydrophobic in the regulatory domain present alignment, yet smaller and hydrophobic in the regulatory domain absent population. By mutating the phenylalanine to a small hydrophobic residue in *Nme*IPMS and *Nme*IPMS K395Term, it may alter the stability of subdomain II and provide further information about the role of this subdomain in conferring catalysis in the presence and absence of a regulatory domain. The converse mutation, a small hydrophobic residue to a large one, could also be made in a protein that does not contain a regulatory domain, such as *Tth*HCS, which has isoleucine present at this position and has been well characterised previously, to observe the difference altering the hydrophobic core of subdomain II makes to the catalytic cycle.

It is obvious that complex interactions are involved in the catalytic cycle of this protein, and an intricate mix of flexibility and restriction is required to facilitate catalysis. The method by which catalysis is maintained through evolution in the presence and absence of a regulatory domain is not entirely clear, but these results provide further information about the mechanism by which AcCoA is recruited to the active site, and the fundamental importance of subdomain II in facilitating that recruitment.

Chapter 5: Modular domain evolution in the IPMS and IPMS-like proteins

5.1 Introduction

A domain can be defined as a functionally independent unit that shows conservation.[159] The same protein domains recur over and over again through evolutionary heritage, appearing in vastly different proteins, performing diverse roles. A considerable number of protein domains have been found both independently, and as part of multi-domain proteins. Examples of these so-called modular domains include the ACT domain, a βαββαβ fold that binds small molecules and can be found through the genomes of a vast array of organisms both as a single domain, and in concert with other domains typically as a regulatory domain of an allosteric protein, and the Src homology 2 (SH2) domain, commonly found in a wide variety of adapter proteins in receptor tyrosine kinase pathways as it binds phosphotyrosine.[160, 161]

Modular domain evolution is critically important in processes such as cell signalling. Peisajovich *et al.*[162] created a domain recombination library of 66 protein variants, mixing regulatory and catalytic domains, from 11 proteins that form part of a yeast mating pathway. The variant proteins were then added individually back to a strain containing the mating pathway, and the dynamics of the pathway were analysed by flow cytometry. The dynamics of the pathway were altered either positively or negatively by the addition of a recombinant variant protein, showing that modular domain evolution is critical to the maintenance and control of cell signalling pathways. Another similar example is the actin cytoskeleton of eukaryotic cells, where its morphology is controlled by guanine exchange factors (GEFs) that activate Rho family GTPases.[163] Dbl family GEFs, that have a conserved Dbl homology (DH) domain but a wide variety of regulatory domains, were interchanged with different regulatory modules that are sensitive to other ligands, the actin cytoskeleton can be altered in predictable ways in response to the new regulatory ligands. This shows how cell signalling pathways can be altered in very wide-ranging ways by simply exchanging the regulatory modules in proteins that form part of the pathway.

Modular domains are also important in the evolution of allosteric enzymes that form part of primary or secondary metabolic pathways. Many allosteric proteins contain multiple domains, and often, the same small-molecule binding domains, such as the aforementioned ACT domain, are

found on many allosteric enzymes with diverse functions. The ACT or ACT-like domain is found in *E. coli* 3-phosphoglycerate dehydrogenase (3-PGDH)[164] as well as *Thermatoga maritima* DH7PS (*Tma*DH7PS)[12], *Campylobacter jejuni* ATP PRTase, aspartate kinase, and chorismate mutase[160], as well as other enzymes typically involved in amino acid metabolism. In these enzymes, the ACT domain functions as a regulatory domain and binds small molecules such as amino acids. Interestingly, the ACT domain can be found in many different configurations in the quaternary structure of various proteins.[165] This modularity was exploited by Cross et al.[35] where an ACT domain from *Tma*DAH7PS was attached to an unregulated DAH7PS from *Pyroccocus furiosus*. The addition of the regulatory domain provided regulation by tyrosine to *Pfu*DAH7PS without adversely affecting catalysis, demonstrating the ease by which an unregulated protein may gain regulation throughout evolution.

Like regulatory domains, catalytic domains can also be modular. One major example is the $(\beta/\alpha)_8$, or TIM, barrel. This domain is found in an enormous number of enzymatic contexts, such as alanine racemase, α-amylase, phosphotriesterase, and dihydropteroate (DHP) synthetase.[70] The barrels can bind a diverse range of cofactors, such as divalent metals or pyridoxal-5'-phosphate, and can catalyse a truly spectacular number of reactions, acting as a lyase, hydrolase, transferase, oxidoreductase, and others.[166] It was originally thought that the diversity of the barrel arose from convergent, rather than divergent evolution, but the similar location of catalytic residues at the C-terminal end of the sheets of the barrel, even though the catalytic functions are incredibly diverse, suggest that, although sequence similarity between enzymes containing the TIM barrel is very low, it probably arose from divergent evolution from a catalytically promiscuous ancestral protein.[70]

This fold also forms the catalytic domain of IPMS, CMS, and HCS, demonstrating the diverse range of substrates that the catalytic barrel can bind even in this small subset of enzymes. Like other examples, such as malyl-CoA lyase which also contains an insertion domain followed by a β-hairpin C-terminal to the barrel, the IPMS/CMS/HCS group of enzymes also contain extensions C-terminal to the barrel.[167] In the case of IPMS and CMS, there is a C-terminal regulatory domain. Interestingly, this regulatory domain appears to be a novel fold as structural searches suggest that the closest similar structure is a double-stranded RNA binding domain.[67]

The origin of the IPMS/CMS regulatory domain is of particular interest as it is a novel fold. It is not present on any characterised protein outside of IPMS and CMS. This could be due to some as-of-yet undiscovered role that is specific to the regulatory domain of these proteins and only this particular fold forming regulatory domain. However, this idea can be disputed with the discovery

that IPMS is functional without a regulatory domain, suggesting that the regulatory domain cannot perform a significant role in catalysis.[2]

As homocitrate synthase does not have a regulatory domain, this suggests that the regulatory domain is a recent addition, after the divergence of the homocitrate synthases. It is presumed that the original, ancestral, promiscuous IPMS/CMS/HCS was not allosterically regulated, as allosteric and other types of regulation is a key step in increasing enzyme specificity throughout evolution.[168] This could also be important to explain why the regulatory domain has not spread throughout the genome – it is a relatively recent novel fold and thus has not had sufficient time to spread through the genome. There may also be genetic factors as to why this regulatory domain has not moved as a modular unit, for example, it may simply not be sufficiently near to a recombination 'hot spot' to allow for movement of the domain as a functional unit.

Moore et al.[113] suggest that 'orphan domains' such as the IPMS/CMS regulatory domain could be recent in evolutionary origin or could be derived from an ancestral protein fold yet sufficient evolutionary distance has occurred so the 'novel fold' and the original ancestral protein are not connected by existing methods of domain association. The authors also suggest that another way in which orphan domains can occur is through mutation of the stop codon and subsequent transcription and translation of formerly non-coding regions of DNA.[113] However, these are typically disordered, whereas the IPMS/CMS regulatory domain has significant secondary structure. Hypothetically, the regulatory domain may formerly have been part of a significantly bigger domain, with a somewhat different fold, but most of the domain, and the original fold, was lost as it was extraneous to the functioning of the enzyme, and thus the 'novel' fold of the ligand-binding portion of the domain remains.

Another argument that could be made about the unique fold of the regulatory domain is that it cannot fold independently of the rest of the protein, and thus is selected against in the genomes of organisms as this may disrupt correct tertiary structure formation, and thus is not found as a modular domain.

As IPMS, existing naturally without a regulatory domain, have been found and characterised, and the regulatory domain has been removed from both IPMS and CMS to produce catalytically active proteins, the regulatory domain cannot play an obligatory role in catalysis, yet, thus far, no IPMS or related protein with a different type of regulatory domain has been characterised. In a review of the ACT domain, Grant[160] mentioned a LeuA protein from *Sulfolobus solfataricus* (*Sso*) that has a putative regulation of amino acids (RAM) domain C-terminal to the subdomains. The RAM

domain is structurally similar to an ACT domain, forming a β-β-α-β fold , but has a different effector binding site.[169]

The ACT domain is a small-molecule binding domain (SMBD) that confers regulation to a protein upon ligand binding to the ACT domain. It is found in numerous different enzymes, often those involved in amino acid biosynthesis such as prephenate dehydrogenase and *E. coli* 3-phosphoglycerate dehydrogenase where the ACT domain binds an allosteric inhibitor to regulate the protein.[164] Like the ACT domain, the RAM domain can be found as a stand-alone small molecule binding domain, or fused to a catalytic domain or a DNA binding domain.[169] One example is found in the thermophilic bacteria *Thermus thermophilus* (*Tth*), where a stand-alone RAM SMBD termed SraA, in the presence of tryptophan, forms a decamer and a complex with anthranilate phosphoribosyltransferase (AnPRT) to provide regulation by tryptophan to AnPRT.[170] It has also been shown that lysine biosynthesis in *Sso* is controlled at a transcriptional level by LysM, a homologue of leucine responsive protein (Lrp), that contains a RAM domain that binds lysine and decreases the affinity of the DNA binding domain for the *lysW* promoter that controls the *lys* operon.[171]

Kumar et al.[69] explored a small group of proteins containing the ACT-like domain noted by Grant et al.[160] further, showing that the proteins containing this structural motif are exclusively found in Archaea, and are likely to be HCS. Additionally, multiple sequence alignments suggested that, as an aspartate residue in the active site that is critical for competitive inhibition by lysine in canonical HCS is absent in the proteins containing the RAM domain, this group of HCS proteins may be allosterically regulated.[69] Attempts were made to characterise the protein from *Sso* were unsuccessful.[69]

To explore the potential for domain modularity in these proteins, the HCS from *Sso* (*Sso*HCS) discussed by Kumar et al.[69] was cloned, expressed, and partially purified. Kinetic activity and inhibition of the partially purified protein were explored. Additionally, domain fusions were made with the *Nme*IPMS catalytic scaffold, where the *Nme*IPMS regulatory domain was replaced with the regulatory domain from *Sso*HCS, and the regulatory domain from *Lin*CMS, to assess whether catalysis could be preserved, and inhibition by a different amino acid could be introduced by swapping the regulatory domain. Several fusions were also made between *Spo*HCS and *Nme*IPMS to investigate the interchangeability of the subdomains between allosterically regulated and competitively inhibited proteins.

5.2 Results

5.2.1 Cloning and purification of *Sso*HCS

The gene and protein sequences of *Sso*HCS were obtained from KEGG (ID: SSO0977).[172-174] The genome sequence of *Sso* had previously been determined, and the gene coding for *Sso*HCS had been annotated as *leuA*-2.[175] The proteome of *Sso* P2, one of the reference strains, has been explored, and the isoelectric point (pI) of *Sso*HCS was determined to be 7.2 while the molecular weight or *Sso*HCS was 50881 Da.[176]

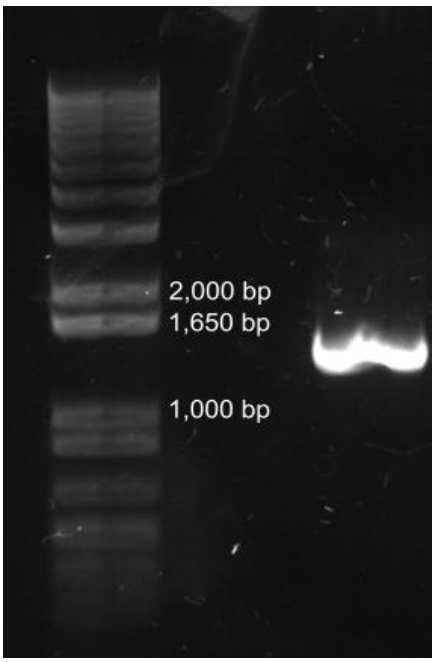


Figure 5.1: Agarose gel of the cloned *Sso*HCS gene (approx. size: 1350 bp)

The gene coding for *Sso*HCS was cloned from *Sso* genomic DNA (Figure 5.1). The gene was then cloned into pET28a, a vector that contains a N-terminal His_6 tag and provides kanamycin resistance, using the InFusion® HD cloning kit (Clontech). The vector contains a thrombin cleavage site between the N-terminal His_6 tag and the gene sequence, and this was replaced with a TEV protease cleavage site. T7 primers were then used to sequence the gene.

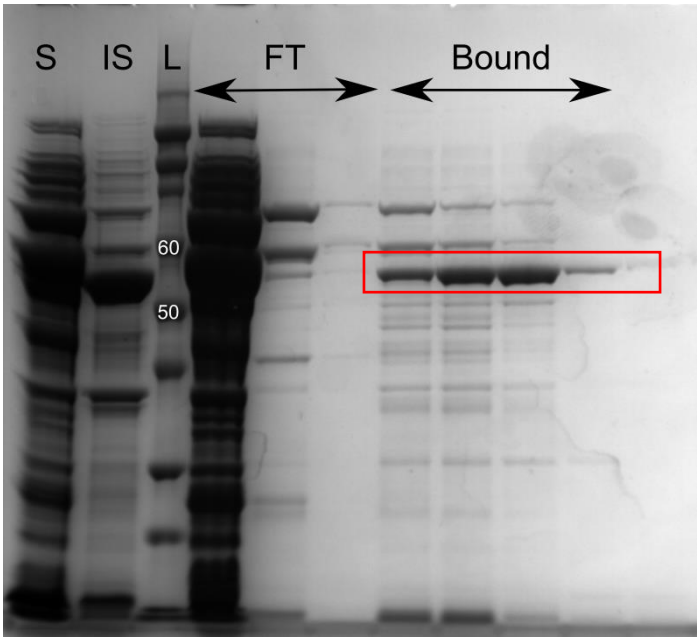


Figure 5.2: IMAC purification of *Sso*HCS. S denotes the soluble fraction, IS the insoluble fraction, L the molecular weight ladder, FT the flow-through from the column, and Bound, the protein that bound to the IMAC column including *Sso*HCS (red box).

The *Sso*HCS construct was insoluble under most conditions tested, including variations in type of buffer, salt concentration, type of salt, addition of additives such as metal ions or detergents, and different lysis methods such as chemical or physical lysis methods. Constructs were also made with GST and MBP solubility tags added using the Gateway system. The protein was not soluble with the GST tag added and showed some solubility, but no activity, with the MBP tag added.

Using a lysis buffer of Tris, pH 8.5 with 100 mM NaCl, marginal solubility was achieved. The protein was partially purified using IMAC, although the low concentration of soluble protein in the lysate allowed for substantial non-specific binding to the IMAC column (Figure 5.2). To improve the purity of the protein would have reduced the already low yield, and further investigation into improving the purification protocol are on-going. Further purification, such as by SEC, was not attempted due to the small amount of protein obtained by this method, and the His_6 tag was not removed.

One way to improve purification may be to utilise the high thermal stability of the *Sso*HCS protein compared to *E. coli* proteins and use heat treatment followed by centrifugation to remove *E. coli* proteins. Ion exchange chromatography, namely cation exchange, could also be investigated instead of using the His_6 affinity tag to purify *Sso*HCS.

5.2.2 Michaelis-Menten kinetics of *Sso*HCS

Kinetic parameters of the partially purified His_6 tagged *Sso*HCS were obtained. Activity was tested at a range of temperatures as *Sso* is a hyperthermophilic species. However, AcCoA is a heat labile substrate and produced substantial background activity at increased temperatures as the chemical couple that reacts with the free thiol of CoA reacted with CoA produced by AcCoA degradation as well as CoA produced by the enzymatic reaction. A temperature of 60°C was selected as this produced the best trade-off between background activity and catalytic activity of the thermophilic protein.

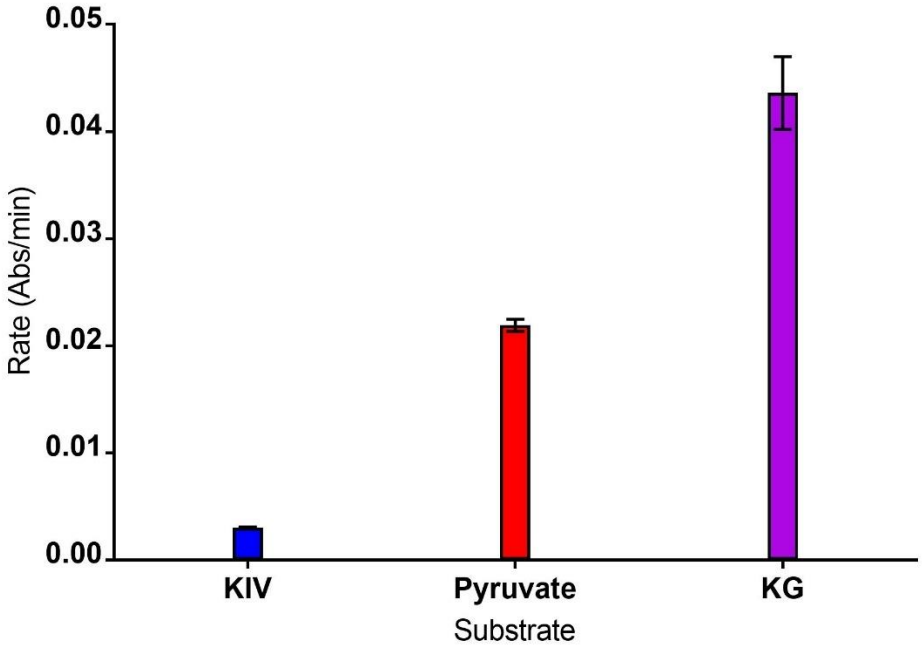


Figure 5.3: The substrate preference of *Sso*HCS. The change in rate was determined using 17 μg of protein per experimental condition. The concentration of ketoglutarate (KG), ketoisovalerate (KIV), and pyruvate was 250 μM and the concentration of AcCoA was 250 μM.

Ketoisovalerate (KIV), pyruvate, and ketoglutarate (KG) were all tested as potential substrates in combination with AcCoA (Figure 5.3). The highest level of activity was seen with ketoglutarate, suggesting that this protein acts primarily as a homocitrate synthase. Additionally, there was some activity observed with pyruvate, but none with ketoisovalerate, suggesting that the active site can accommodate other substrates, as seen in other members of the broad protein family. However, due to the low amount of soluble protein obtained, further exploration of other alternative substrates was not pursued.

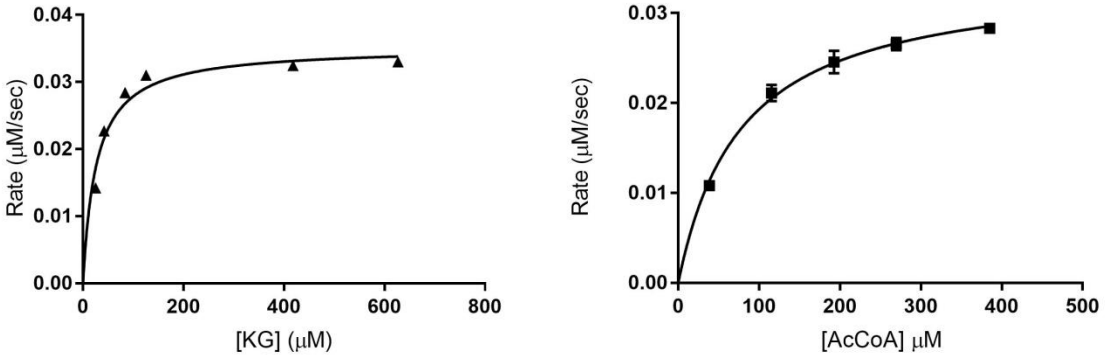


Figure 5.4: Plot of the initial rate of *Sso*HCS in response to a change in substrate. The plot for increasing KG is on the left and increasing AcCoA is on the right. When KG was varied, AcCoA was held at 500 μM, and when AcCoA was varied, KG was held at 250 μM.

Table 5.1: Kinetic parameters for *Sso*HCS

	K_m (KG, μM)	K_m (AcCoA, μM)	k_{cat} (s^{-1})
*Sso*HCS	27 ± 5	78 ± 7	0.160

Preliminary characterisation of the protein was performed using the substrates ketoglutarate and AcCoA in HEPES buffer at pH 7.5 with 20 mM $MgCl_2$ and 20 mM KCl. Other buffers were tested, at a range of pH levels, but this buffer produced the most consistent activity. High pH can cause the reaction between DTP and CoA to become rate-limiting, so a larger range of buffers could be tested in the future using the direct as opposed to the indirect assay, where instead of using a chemical couple (DTP) that reacts with CoA to form a thio-pyridine that can be detected at 324nm, the loss of AcCoA at 232 nm can be directly measured. This was not performed due to time restraints.

5.2.3 Testing of inhibitors

Inhibition by three amino acids was tested (Figure 5.5). Due to the low amount of protein obtained, the activity at saturated substrate concentrations in the presence of 1 mM of L-lysine, L-leucine, or L-isoleucine, was tested and compared to the activity at the same concentration of substrates in the absence of inhibitor. The activity decreased to approximately half the maximal activity in the presence of 1 mM L-lysine, but not in the presence of either L-leucine or L-isoleucine. This suggests that the protein is inhibited only by L-lysine. In combination with evidence from multiple sequence alignments that the *Sso*HCS protein lacks a key aspartate residue in the active site to allow for competitive inhibition by lysine, it is plausible that this is the first evidence of an allosterically inhibited HCS.[69]

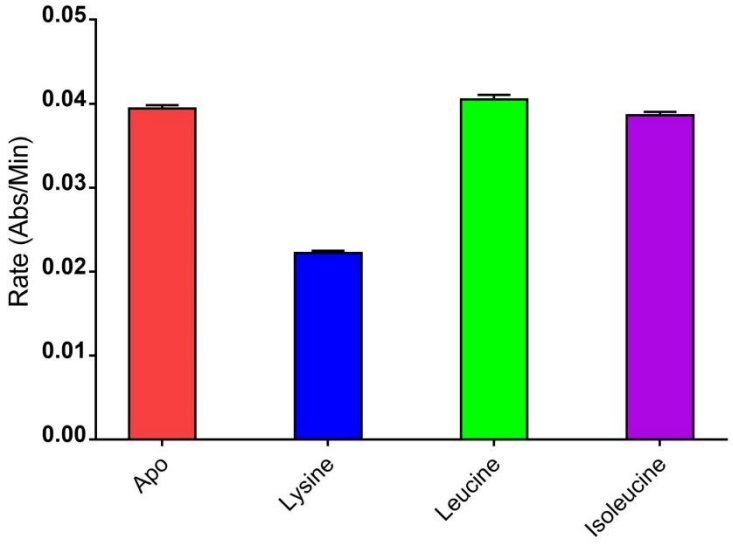


Figure 5.5: The response of *Sso*HCS to potential inhibitors. The concentration of both substrates was held at 500 μM when an inhibitor was tested. The concentration of all inhibitors tested was 1 mM.

5.2.4 Summary and Discussion

The putative HCS from *Sso* was cloned and partially purified. The problems with solubility outlined by Kumar et al.[69] were partially overcome by the screening of different buffers, although further buffer screening may be necessary to obtain higher levels of protein, or a different growth strategy, such as changing the media in which *E. coli* is grown to increase the cell concentration, may be of use. Additionally, a method of purification such as heat treatment, as the protein is thermostable, may be more appropriate than purification using a His_6 tag. Other additives such as different detergents may also improve solubility. The purified protein displayed maximal activity at increased temperatures, as expected of a protein from a thermophilic organism. The protein also showed inhibition by lysine, suggesting that it may be allosterically regulated. Further work is needed to establish the range of substrates that the protein can accommodate and to explore the mode of inhibition by lysine.

5.2.5 Fusion proteins

To further explore the potential for modularity in IPMS and related proteins, several fusion proteins were made to investigate whether the subdomains of *Nme*IPMS and *Spo*HCS were interchangeable. A fusion was also made between *Spo*HCS and the regulatory domain of *Nme*IPMS, to investigate whether a HCS, which naturally does not have a regulatory domain, could still perform catalysis with a regulatory domain attached to its C-terminus.

Additionally, two fusions were made where the regulatory domain of *Lin*CMS or the putative regulatory domain of *Sso*HCS was fused to the end of subdomain II of *Nme*IPMS. This was done to explore whether the regulatory domains of IPMS and related proteins were interchangeable, and also to assess whether inhibition by isoleucine or lysine respectively could be transferred between the proteins by fusion of the regulatory domain.

5.2.5.1 Construction of fusion proteins

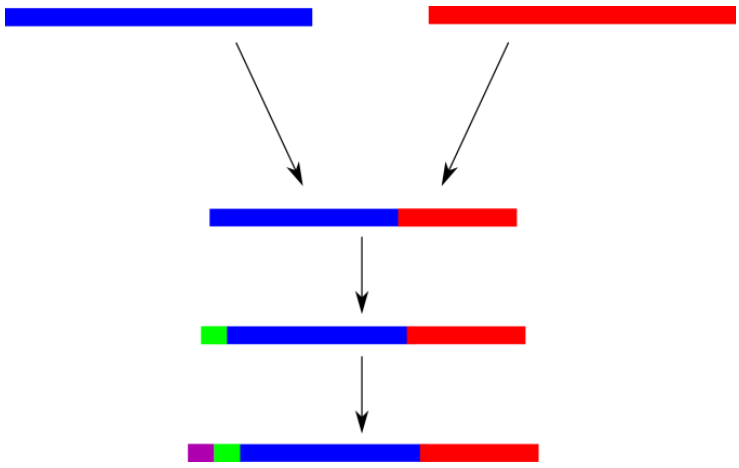


Figure 5.6: The strategy used to fuse genes together in the creation of the fusion construct. The red and blue lines denote the two genes, while the purple and green lines denote the His_6 tag and the TEV protease site that were also added.

A series of pairwise, structural, and multiple sequence alignments was used to assess where to fuse the two fragments of the genes encoding the full-length protein together. The *Nme*IPMS/*Spo*HCS fusions were made at positions to maintain the entirety of both subdomains, thus preserving subdomain II as a single entity, as this is known to be essential for catalysis. All of the fusion proteins were constructed using multi-step PCR and the plasmid maps for the *Spo*HCS and *Sso*HCS constructs are located in Appendix IV. The protein sequences for the fusion constructs are located in Appendix IV. The *Spo*HCS construct was obtained from GeneArt and was sub-cloned into

pET28a for protein expression. The *Lin*CMS construct was also a synthetic gene and was also sub-cloned into pET28a. The *Nme*IPMS and *Sso*HCS constructs were described above. A table of the fusion proteins that were constructed is shown in Table 5.2.

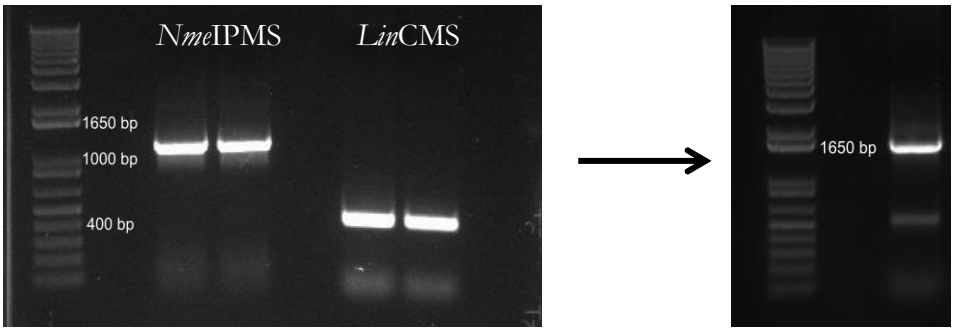


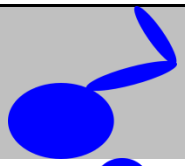
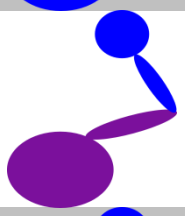
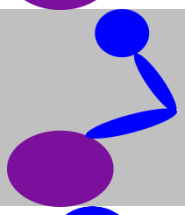
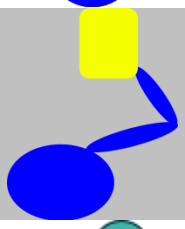
Figure 5.7: An example of the first two PCR stages, where regions of genes of interest are amplified and then fused using the overlapping region designed into the primer. These gene fragments were generated during the construction of the *Nme*IPMS-*Lin*CMS fusion and show the gene fragment associated with *Nme*IPMS catalytic and subdomains in the lanes labelled *Nme*IPMS, and the gene fragment that codes for the regulatory domain of *Lin*CMS in the lanes labelled *Lin*CMS. The gene fragment in the right-hand image shows the fused gene construct where the gene fragment coding for the *Lin*CMS regulatory domain has been fused to the gene fragment coding for the *Nme*IPMS catalytic domain and subdomains using overlapping primers.

The cloning strategy used is described in Figure 5.6. The part of each gene of interest was amplified using PCR, then gel purified. An example of this is shown in Figure 5.7. In the primer, a region of the other gene was included on the relevant end to allow for overlap of the two constructs in the second PCR step. A further two rounds of PCR were used to add an N-terminal His tag and a TEV protease site to the construct. InFusion cloning was then used to move the fused gene into pET21a for expression.

5.2.5.2 Fusions of *Nme*IPMS and *Spo*HCS

These fusions were constructed to assess the modularity of subdomain II in particular by determining whether a complete subdomain II from a different structural design, i.e. with or without a regulatory domain respectively, could be sufficient for catalysis.

Table 5.2: The fusion constructs, showing the regions of each protein that they contain. The residue numbers given are the residue number of that particular protein i.e. the residue numbers given for *Nme*IPMS are the residue numbers from *Nme*IPMS, the residue numbers for *Lin*CMS are from *Lin*CMS etc. In the schematic diagrams of the single chain of the protein, parts of *Spo*HCS are shown in purple, parts of *Nme*IPMS are shown in blue, parts of *Lin*CMS are shown in teal, and parts of *Sso*HCS are shown in yellow.

Name	Part of *Nme*IPMS	Part of *Spo*HCS	Part of *Lin*CMS	Part of *Sso*HCS	Schematic
*Nme*IPMS K395	Catalytic domain and both subdomains (Residues 1-395)	-	-	-	
*Spo*HCS$_{Cat-SI}$ – *Nme*IPMS $_{SII-Reg}$	Subdomain II and regulatory domain (Residues 330-517)	Catalytic domain and subdomain I (Residues 1-351)	-	-	
*Spo*HCS$_{cat}$ – *Nme*IPMS$_{SDs-Reg}$	Regulatory domain (Residues 395-517)	Catalytic domain and both subdomains (Residues 1-413)	-	-	
*Nme*IPMS$_{Reg}$	Regulatory domain (Residues 395-517)	-	-	-	
*Nme*IPMS$_{Cat-SDs}$ – *Sso*HCS$_{Reg}$	Catalytic domain and both subdomains (Residues 1-394)	-	-	(Putative) regulatory domain (Residues 399-461)	
*Nme*IPMS$_{Cat-SDs}$ – *Lin*CMS$_{Reg}$	Catalytic domain and subdomains (Residues 1-388)	-	Regulatory domain (Residues 388-516)	-	

5.2.5.2.1 The fusion of the *Spo*HCS catalytic domain and subdomain I with the *Nme*IPMS subdomain II and regulatory domain (*Spo*HCS$_{Cat-SI}$ – *Nme*IPMS$_{SII-Reg}$)

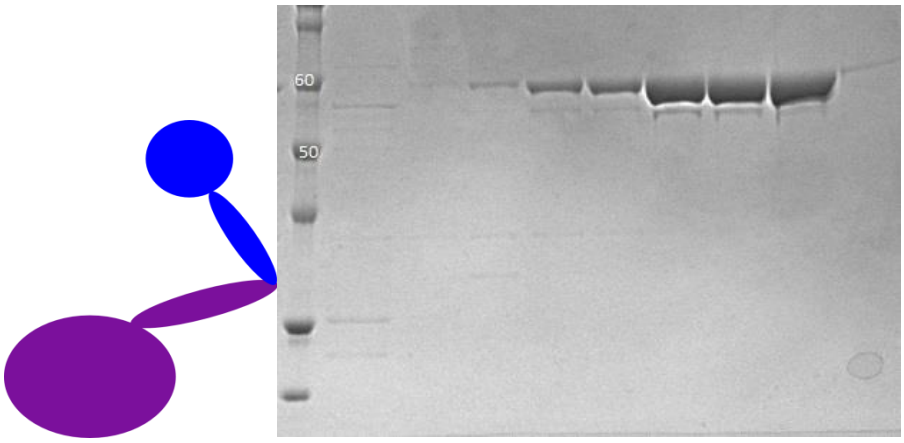


Figure 5.8: A schematic diagram of *Spo*HCS$_{Cat-SI}$ – *Nme*IPMS$_{SII-Reg}$. (left) The catalytic barrel and subdomain I from *Spo*HCS are in purple, while subdomain II and the regulatory domain from *Nme*IPMS are shown in blue. The HisTrap purification of *Spo*HCS$_{Cat-SI}$ – *Nme*IPMS$_{SII-Reg}$ (right).

The *Spo*HCS$_{Cat-SI}$ – *Nme*IPMS$_{SII-Reg}$ construct was made by fusing *Spo*HCS to residue Gly351 (*Spo*HCS numbering) to *Nme*IPMS from residue Leu330 (*Nme*IPMS numbering). This encompasses *Spo*HCS to the C-terminal end of subdomain I, and *Nme*IPMS from the N-terminal end of subdomain II to the end of the regulatory domain. The aim of this fusion was to determine whether an HCS with no regulatory domain could be catalytically active with a regulatory domain attached. Subdomain II from *Nme*IPMS was included in the fusion to allow for potentially important connections between subdomain II and the regulatory domain to be maintained in the overall structure.

Following the PCR steps detailed above, the purified plasmid was sequenced, and this showed that the correct fusion had been made and there were no other errors in the sequence. A HisTrap purification of this fusion protein was performed after it was determined that the protein was soluble and approximately the correct size (Figure 5.8). Further purification of the protein was not performed as only a small amount was obtained from the HisTrap purification step. There was some contamination in the HisTrap purification that could not be adequately resolved to determine whether the protein was correctly folded by circular dichroism nor whether the mass was consistent with the estimated mass using mass spectrometry. The protein also displayed no catalytic activity upon kinetic assay using ketoglutarate and AcCoA as substrates. The maximum concentration of protein tested was 0.075 mg/mL.

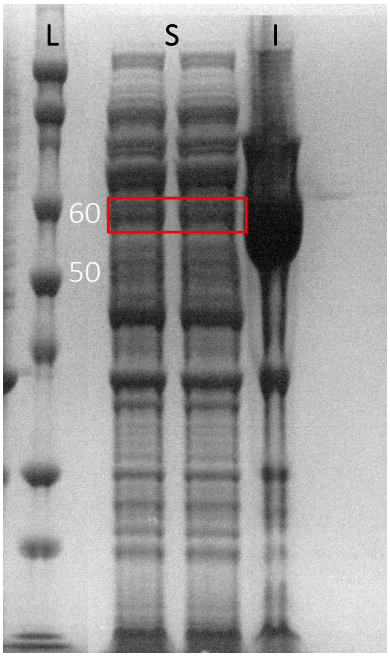


Figure 5.9: The insolubility of the *Spo*HCS$_{cat\text{-}SDs}$ – *Nme*IPMS$_{Reg}$ fusion. This gel image shows the soluble (S) and insoluble (I) fractions after lysis of the *Spo*HCS$_{cat}$ – *Nme*IPMS$_{SDs\text{-}Reg}$ fusion protein by sonication in a buffer containing 50 mM potassium phosphate pH 8, 300 mM KCl, 20 mM imidazole. L denotes the ladder. A significant amount of insoluble protein is observed in the insoluble fraction, although a small amount (red box) was soluble.

A further fusion, where the regulatory domain from *Nme*IPMS was fused to the C-terminal end of subdomain II of *Spo*HCS was also constructed (*Spo*HCS$_{cat\text{-}SDs}$ – *Nme*IPMS$_{Reg}$). However, this fusion was substantially less stable than the *Spo*HCS$_{Cat\text{-}SI}$ – *Nme*IPMS$_{SII\text{-}Reg}$ fusion, and showed significant precipitation, and a very low yield, when an attempt was made to purify it using IMAC. When the partially purified protein was thawed after flash-freezing and storage in a -80°C freezer, more precipitation occurred, and no activity was observed by kinetic assay. Further investigation of buffer conditions that may stabilise this protein to enable purification and characterisation are on-going.

5.2.5.3 Fusions of *Nme*IPMS and *Lin*CMS, and *Nme*IPMS and *Sso*HCS

To further explore the plausibility of the modularity of the regulatory domain from these proteins, the regulatory domain was purified independently of the rest of the protein, and two fusion proteins were constructed where the regulatory domain from *Lin*CMS, or the putative regulatory domain from *Sso*HCS, were fused to the catalytic unit, that is the catalytic domain and subdomains, of *Nme*IPMS to observe whether catalysis could be preserved, and whether allostery by the respective allosteric inhibitor could be introduced.

5.2.5.3.1 The *Nme*IPMS regulatory domain (*Nme*IPMS$_{reg}$)

The *Nme*IPMS regulatory domain was cloned independently of the rest of the protein. The protein was soluble, but purification of the regulatory domain by the N-terminal His$_6$ tag was difficult as there was a large amount of contamination in the purification that could not be adequately resolved (Figure 5.10). The regulatory domain appeared to be able to be purified as soluble protein, although whether the domain is properly folded, and whether it forms the correct oligomeric state and can bind leucine is yet to be established.

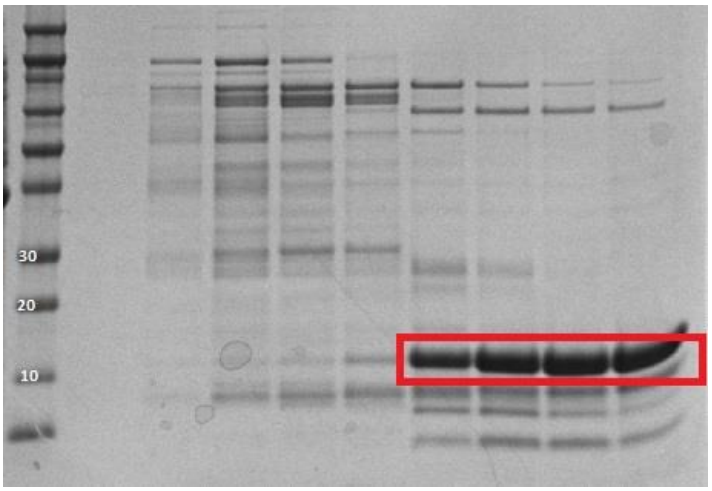


Figure 5.10: HisTrap purification of the isolated *Nme*IPMS regulatory domain. The red box indicates the protein of approximately the size of the *Nme*IPMS regulatory domain. HisTrap purification was performed using an elution gradient from 0 to 0.5 mM of imidazole. This protein was eluted from the HisTrap in the latter part of the gradient.

5.2.5.3.2 The *Nme*IPMS-*Sso*HCS fusion (*Nme*IPMS$_{Cat\text{-}SDs}$ – *Sso*HCS$_{Reg}$)

This fusion was made to explore the putative *Sso*HCS regulatory domain by fusion to the *Nme*IPMS catalytic unit to see whether it was able to provide allosteric regulation by L-lysine to *Nme*IPMS, and potentially, to observe whether lysine could bind this domain in this context as producing sufficient soluble *Sso*HCS for protein-intensive techniques such as ITC proved difficult. Additionally, this fusion was constructed to determine whether *Nme*IPMS was catalytically active with a different type of regulatory domain fused to the C-terminus.

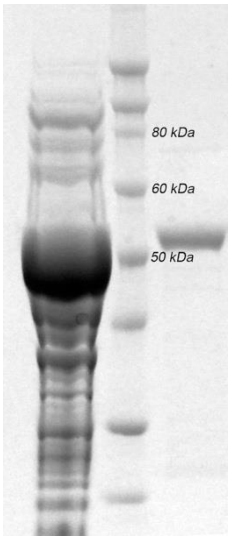


Figure 5.11: Purification of the *Nme*IPMS$_{Cat\text{-}SDs}$–*Sso*HCS$_{Reg}$ fusion protein. The first lane shows the soluble fraction of the lysate, showing large amount of soluble fusion protein (red box). The second lane contains a molecular weight ladder. The third lane shows the fusion protein after HisTrap and SEC purification, showing that the fusion protein is relatively pure and is approximately the correct size that was calculated as 51700 kDa.

As described above, this fusion was made using PCR where an overlap region with *Sso*HCS was introduced at the C-terminus of subdomain II of *Nme*IPMS, and the N-terminus of the putative regulatory domain of *Sso*HCS. As the absolute structure of *Sso*HCS is not known, the regulatory domain was defined by the residues identified by Pfam[114] as an Asnc/Lrp ligand binding domain otherwise known as a RAM domain. The regulatory domain in this context was thus from residue 399 (*Sso*HCS numbering) to the C-terminal end of the protein. The fusion point on *Nme*IPMS was made at residue Tyr394, as the active truncation of *Nme*IPMS was made at this position and therefore encompassed all of subdomain II. The protein sequence of this and the other fusions constructed are located in Appendix IV.

The protein was soluble, and purification by IMAC followed by SEC was performed. As discussed previously, removal of the His_6 tag led to a substantial decrease in the amount of active protein obtained, so the His_6 tag was also not removed from the fusion protein constructs.

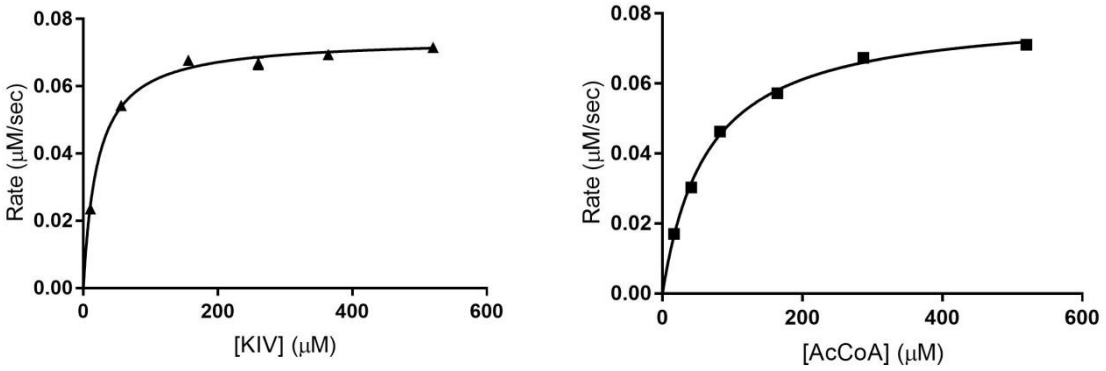


Figure 5.12: Michaelis-Menten kinetic data for the *Nme*IPMS$_{Cat\text{-}SDs}$–*Sso*HCS$_{Reg}$ fusion protein. The concentration of KIV was varied in the left plot and the concentration of AcCoA was varied in the right plot. The other substrate was held at a saturating concentration of 250 μM (KIV) and 300 μM (AcCoA)

Michaelis-Menten kinetics were obtained for the purified *Nme*IPMS$_{Cat\text{-}SDs}$–*Sso*HCS$_{Reg}$ construct, using ketoisovalerate and AcCoA as substrates (Figure 5.12). The protein is catalytically active and has a K_m for KIV similar to that of wild type *Nme*IPMS (Table 5.3). The K_m for AcCoA of the *Nme*IPMS$_{Cat\text{-}SDs}$–*Sso*HCS$_{Reg}$ fusion protein has increased compared to the wild type protein, to approximately a similar degree as the K_m for AcCoA for the *Nme*IPMS K395Term truncated protein. This suggests that the fusion of the *Sso*HCS regulatory domain may somewhat adversely affect the interaction of subdomain II with AcCoA. Connections subdomain II may make with the regulatory domain in the wild-type protein have been broken in this construct, and the lack of these connections may increase the conformations that the subdomains may form, leading to a decrease in the ability of the subdomains to recruit AcCoA to the active site.

The K_m for KIV has decreased in the fusion protein compared to the wild type protein from 36 μM to 21 μM. This decrease suggests that the addition of the regulatory domain from a more thermostable protein has increased the ability of the protein to interact with KIV.

The k_{cat} of the fusion protein has decreased substantially compared to the wild type protein, although as mentioned above, the k_{cat} in these proteins can be variable due to the proportion of active to inactive protein as the proteins lack stability.

The *Nme*IPMS$_{Cat\text{-}SDs}$–*Sso*HCS$_{Reg}$ protein does not show inhibition by L-lysine, L-leucine, or L-isoleucine up to concentrations of 10 mM, suggesting that inhibition via the regulatory domain has

not been introduced. Differential scanning fluorimetry (DSF) was used to investigate whether there was evidence of ligand binding. Typically, if thermal stabilisation is observed by DSF in the presence of a ligand, it suggests that the ligand is binding the protein, although the absence of thermal stabilisation does not preclude ligand binding. Thermal stabilisation by ligand binding was was demonstrated in Chapter 4.1.3 by L-leucine stabilisation of the wild type *Nme*IPMS protein.

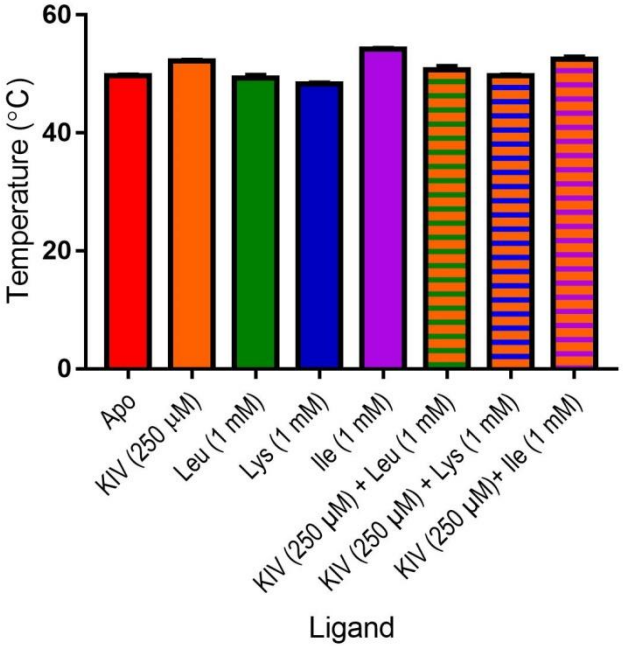


Figure 5.13: The melting temperature of *Nme*IPMS$_{Cat\text{-}SD_3}$–*Sso*HCS$_{Reg}$ as determined by DSF.

The apo protein had a higher T_m than the *Nme*IPMS truncation or the *Nme*IPMS wild type protein. Stabilisation by leucine seen in the wild type *Nme*IPMS truncation was not observed in the *NmeSso* fusion as expected as the *Sso*HCS regulatory domain has been posited to bind lysine, although no stabilisation is seen with lysine either. There is a small amount of stabilisation seen when isoleucine is added, although the T_m is only 2 °C higher than that of the apo protein. This slight increase may be significant, but further examination such as by ITC would be required to investigate ligand binding, particularly isoleucine interaction causing stabilisation observed in the thermal shift assay.

Table 5.3: Kinetic parameters for *Nme*IPMS wild-type, the truncated *Nme*IPMS, and two fusion proteins

Name	K_m (AcCoA, μM)	K_m (KIV, μM)	k_{cat} (s^{-1})
*Nme*IPMS wild type	36 ± 3	35 ± 3	7.2 ± 0.1
*Nme*IPMS K395Term	80 ± 7	30 ± 3	4.1 ± 0.1
*Nme*IPMS$_{Cat-SDs}$ – *Lin*CMS$_{Reg}$	50 ± 5	31 ± 4	1.5 ± 0.1
*Nme*IPMS$_{Cat-SDs}$ – *Sso*HCS$_{Reg}$	64 ± 3	21 ± 1	1.6 ± 0.1

This fusion, although not allosterically regulated, does suggest that a different type of regulatory domain can be fused to the C-terminus of subdomain II of *Nme*IPMS and not significantly adversely affect catalysis. However, allostery was not preserved in this fusion.

5.2.5.3.3 The *Nme*IPMS – *Lin*CMS fusion (*Nme*IPMS$_{Cat-SDs}$ –*Lin*CMS$_{Reg}$)

Another fusion, between the catalytic unit of *Nme*IPMS and the regulatory domain of *Lin*CMS, was thus made to determine whether a canonically structured regulatory domain for this type of protein could be switched between proteins and confer both catalysis and allostery.

This fusion was made to encompass the catalytic unit of *Nme*IPMS, encompassing the protein up to residue Ser388 (*Nme*IPMS numbering) and the regulatory domain of *Lin*CMS from residue Gly388 (*Lin*CMS numbering), and was constructed as described above. *Lin*CMS was chosen as the source of the regulatory domain as, although the two proteins are phylogenetically distant, the structure of the *Lin*CMS regulatory domain has been solved.

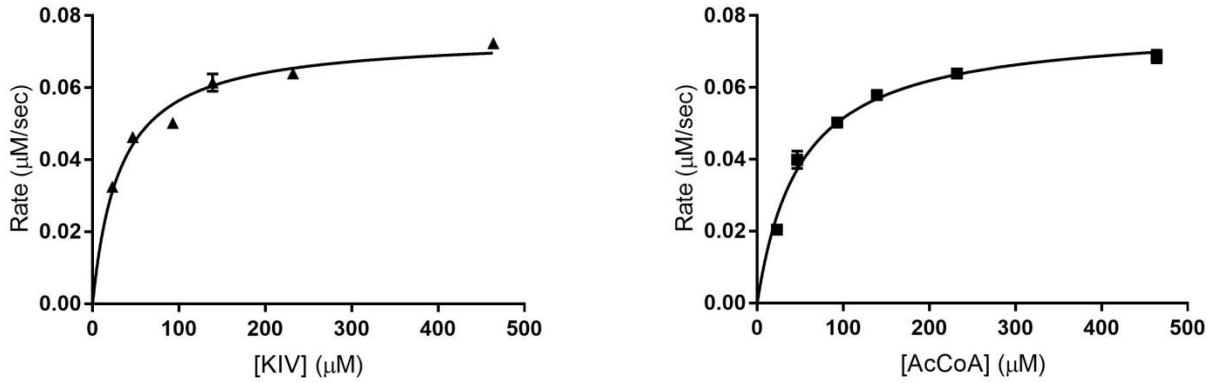


Figure 5.14: Kinetic activity of the *Nme*IPMS$_{Cat\text{-}SDs}$–*Lin*CMS$_{Reg}$ fusion. In the left plot, AcCoA was held at 250 μM, and in the right plot, KIV was held at 250 μM, while the other substrate was varied.

The fusion is catalytically active, and has kinetic parameters similar to those of the wild type protein, aside from an increase in K_m for AcCoA and a decrease in k_{cat} (Table 5.3, Figure 5.14). The K_m for AcCoA of this fusion is comparable to the K_m of the truncated *Nme*IPMS and the *Nme*IPMS$_{Cat\text{-}SDs}$ –*Sso*HCS$_{Reg}$ fusion, showing a 1.4-fold increase. Unlike the *Nme*IPMS$_{Cat\text{-}SDs}$ – *Sso*HCS$_{Reg}$ fusion, the addition of the *Lin*CMS regulatory domain did not cause a decrease in the K_m for KIV.

As with *Nme*IPMS$_{Cat\text{-}SDs}$–*Sso*HCS$_{Reg}$, this fusion did not show regulation by L-leucine, L-isoleucine, or L-lysine. Ligand concentrations of the amino acids tested varied from 0 to 10 mM. Additionally, assays were performed, with L-lysine, L-leucine, and L-isoleucine present at varying concentrations (0 – 5 mM), at different concentrations of substrate ranging from low concentrations of either or both substrates (a concentration of 25 μM for AcCoA and/or KIV) to high concentrations of either or both substrates (a concentration of 150 μM for AcCoA and/or KIV). No inhibition was seen under any conditions.

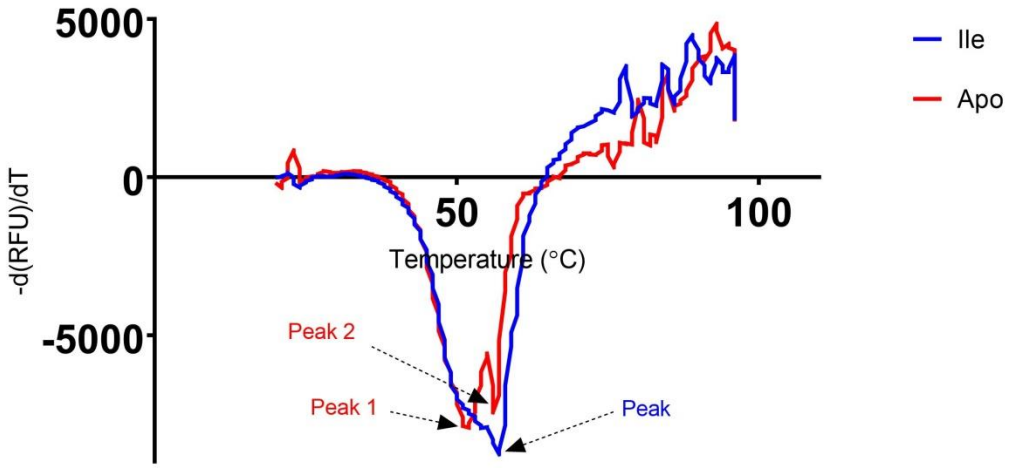


Figure 5.15: DSF denaturation of *Nme*IPMS$_{Cat\text{-}SDs}$–*Lin*CMS$_{Reg}$ without and with 1 mM Ile. The negative of the first derivative is shown in red (apo) and blue (1 mM Ile). This demonstrates the double peak seen in the apo but not the isoleucine-containing condition.

DSF was also used with this fusion to investigate potential ligand binding to the fusion protein (Figure 5.16). *Nme*IPMS$_{Cat\text{-}SDs}$ –*Lin*CMS$_{Reg}$, as with other fusions described above, showed a substantial loss of stability compared to the wild type protein, and precipitated readily upon concentration, meaning that ITC was not a viable technique for exploring ligand binding under these conditions. Two peaks are visible in the graph of the thermal melt in the apo condition, one at 51°C and one at 56°C (Figure 5.15). Under conditions containing isoleucine, the profile had changed with only one main peak observed at 56°C. This suggests that addition of isoleucine has some effect on the stability of the protein.

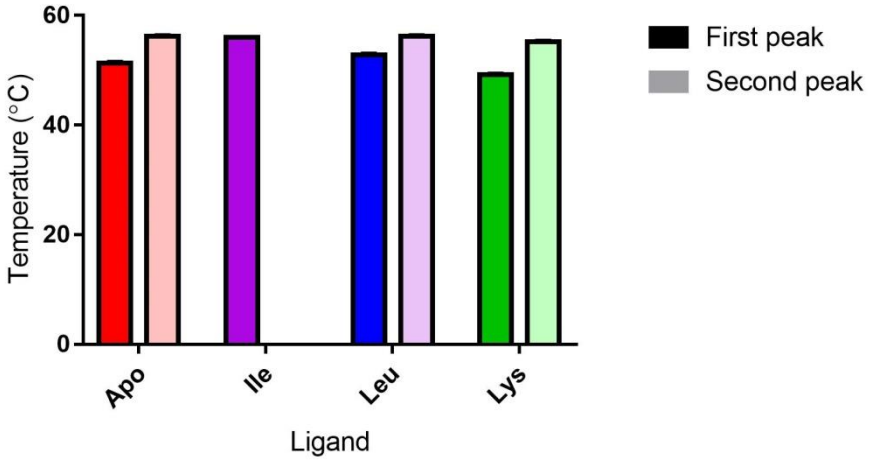


Figure 5.16: Melting temperature for the *Nme*IPMS$_{Cat\text{-}SDs}$–*Lin*CMS$_{Reg}$ fusion as determined by DSF. The temperature of the first peak in each condition is shown in a dark colour and the temperature of the second peak is shown in a light colour. The concentration of each ligand used was 1 mM.

5.3 Discussion

These results show that the point at which a fusion is made is critically important to the maintenance of activity. The fusions made between *Nme*IPMS and *Spo*HCS did not show catalytic activity upon purification, suggesting that a complete catalytic unit, and the interactions that this unit forms within itself; is critical for catalysis. However, a different fusion point, or subsequent mutation, may improve stability in the case of the *Spo*HCS$_{cat}$ – *Nme*IPMS$_{SDs\text{-}Reg}$ fusion or allow for catalytic activity to be maintained in the other *Nme*IPMS-*Spo*HCS fusions. The suitability of other fusion points may be investigated using large MSAs of *Nme*IPMS-like IPMS and *Spo*HCS-like HCS to assess regions of particular conservation in both large groups and when the two groups are aligned together. Another potential area to investigate would be to utilise a HCS from the bacterial HCS as these HCS are phylogenetically closer to *Nme*IPMS than *Spo*HCS are. As the bacterial HCS are from organisms in the *Thermus-Deinococcus* group, they are from extremophilic organisms. The fusion of an enzyme from a mesophilic organism (e.g. *Nme*IPMS) to one from an extremophilic organism (e.g. *Tth*HCS) presents additional complications. Using an extremophilic IPMS, e.g. *Mja*IPMS to create fusions with an extremophilic HCS may also be a plausible solution.

As discussed in Chapter 2, the allosteric network that controls subdomain II and catalytic activity appears to be phylogenetic in basis as opposed to conserved in all proteins of a particular structural type. For example, *Nme*IPMS and *Mtu*IPMS catalyse the same reaction with the same overall structure but appear to have different allosteric networks that transfer the signal from the allosteric ligand binding site to the active site. This difference in network suggests that the transfer of allostery by domain fusion is difficult if the connections to the active site are broken, and, as the results in Chapter 2 show, single point mutations are sufficient to abolish allosteric regulation. However, the transfer of the allosteric signal could potentially be re-established in the fusion protein by subsequent mutations if the allosteric network in the subdomains is not disrupted.

The positive charge of Arg470, located in the regulatory domain, is lost in the *Nme*IPMS$_{SDs}$-*Lin*CMS$_{Reg}$ fusion as there is not a comparable residue on the bottom of the *Lin*CMS regulatory domain. An Arg470Ala mutation in *Nme*IPMS abolished leucine sensitivity, suggesting that this residue is crucial for the transmission of the allosteric signal to the active site. One way to assess whether this is the only factor in the lack of response to the allosteric ligand would be to mutate the comparable residue in the *Lin*CMS domain in the fusion to arginine.

Additionally, the *Lin*CMS is considerably more phylogenetically distinct from *Nme*IPMS than *Mja*CMS is, so the transfer of allostery by the transfer of a domain may be more accessible if the *Mja*CMS regulatory domain is transferred onto the *Nme*IPMS catalytic scaffold. These results suggest that, although catalytic activity can be readily preserved even when a foreign domain is transferred onto the *Nme*IPMS catalytic scaffold, the preservation of allosteric activity is more difficult.

Chapter 6: Discussion

Determining how a protein fluctuates in a living cell is an immense task, although a deeper understanding of how proteins move, how that movement can be affected by allosteric regulation, and the evolutionary processes that have contributed to that movement, are crucially important to our understanding of proteins, their evolution, and ultimately, to our ability to design more effective antibiotics.

In this study, a group of related enzymes have been explored as they share an interesting structural homology, namely a $(\alpha\beta)_8$ barrel with a C-terminal extension, subdomains I and II, that is critical for catalysis. How the subdomains, and specifically how their movement, contribute to the catalytic activity even though they are some distance from the active site, has been of particular interest. A variety of techniques has been utilised to explore the role of the subdomains, how the enzymes have evolved to mediate catalysis both with and without the structural burden of a regulatory domain at the C-terminus of subdomain II, and how the allosteric signal can be transferred to the active site in the absence of an obvious conformational change.

6.1 Residue networks in the subdomains facilitate catalysis in regulatory domain-present and regulatory domain-absent structural populations

In extant IPMS and related enzymes, it has been shown that removal of a regulatory domain, as long as subdomain II remains intact, does not abolish catalysis. This was shown in *Nme*IPMS, in which the information obtained from a previous truncation had suggested that removal of a regulatory domain abolished catalysis although this truncation encompassed only part of subdomain II.[1, 158] There is a catalytic penalty to the removal of the regulatory domain, an increase particularly in the K_m for AcCoA, which suggests that there is a change in the dynamics of the subdomains that recruit AcCoA in the absence of a regulatory domain restraining the conformations that the subdomains can adopt.

In Chapter 3, covariance analyses were used to determine networks of coevolved residues in populations of sequences that possess a regulatory domain and that do not have a regulatory domain. These networks, spanning the catalytic domain and the subdomains, were different in the two populations, suggesting a mechanism by which the subdomains, that appear to have a similar structure in the presence and absence of a regulatory domain, can maintain sufficient freedom in

the presence of a regulatory domain, and sufficient restraint in the absence of one, to confer catalysis. As the truncated form of *Nme*IPMS still contained a network of residues that had evolved to bear the burden of a regulatory domain, there was a catalytic penalty to the removal of the regulatory domain that may have its roots in the increased freedom of the subdomains.

As modern enzymes, such as *Nme*IPMS, can still be catalytically active even with a regulatory domain removed, it suggests that the ancestral protein did not have a regulatory domain, as, if the ancestral protein had a regulatory domain, the network of residues that have evolved to allow catalysis particularly in the subdomains may have encompassed the regulatory domain. Removal of the regulatory domain in modern proteins would thereby be less likely to result in a catalytically active truncated protein. Additionally, Liang et al.[177] suggest that it is likely that allosteric proteins arise from non-allosteric proteins that can catalyse the same reaction. Additionally, domain fusions, where a regulatory domain has been transplanted from one enzyme to another, has been shown to confer allosteric regulation to unregulated proteins.[35, 178]

In Chapter 3, the covariance analyses determined different networks of residues in the proteins that had regulatory domains compared to the proteins that did not, although these networks both included the subdomains and were in similar regions such as the hydrophobic interior of the three-helix bundle of subdomain II. The Lys332Ala mutation made in the linker region of *Nme*IPMS showed a 25-fold increase in K_m for AcCoA showing that the linker between subdomain I and subdomain II is important for catalysis due to its role in the recruitment of AcCoA. As networks identified by covariance analyses include the subdomains, it suggests that these networks are important for the control of the dynamics of the subdomains to facilitate recruitment of AcCoA and allow for catalysis. These networks also show one way these proteins have compensated for the burden of a regulatory domain, as the networks of residues in the regulatory domain present and regulatory domain absent populations are different, suggesting a mechanism by which the dynamics of the subdomains have been altered.

In Chapter 5, it was shown that fusions between *Nme*IPMS and *Spo*HCS were not catalytically active. This suggests that, potentially, interactions between the subdomains, such as in the networks determined in Chapter 3, are important for catalysis and these were disrupted when fusions were made between the subdomains, leading to inactive protein. It also suggests that the catalytic unit in its entirety is the catalytic domain and the subdomains. However, the fusion proteins also showed limited stability which made purification difficult and sufficient amounts of purified fusion protein has not been obtained to allow for circular dichroism to assess whether these proteins are properly folded. With taxonomy in mind, future work could include fusions

between HCS from *Thermus* or *Deinococcus* species as these appear to share a closer phylogenetic relationship to *Nme*IPMS, although evolution towards extreme environments may have altered networks responsible for the control of the subdomains and catalytic activity.

In summary, these results show that a network of residues that span from the catalytic domain to the subdomain are important for catalysis in these proteins both when a regulatory domain is present and when it is absent. This network appears to be important for maintaining the dynamics of the subdomains, allowing flexibility in the presence of a restraining regulatory domain and stability in the absence of one.

6.2 Multiple gene duplication and horizontal gene transfer events have led to the modern taxonomic distribution of catalytic diversity

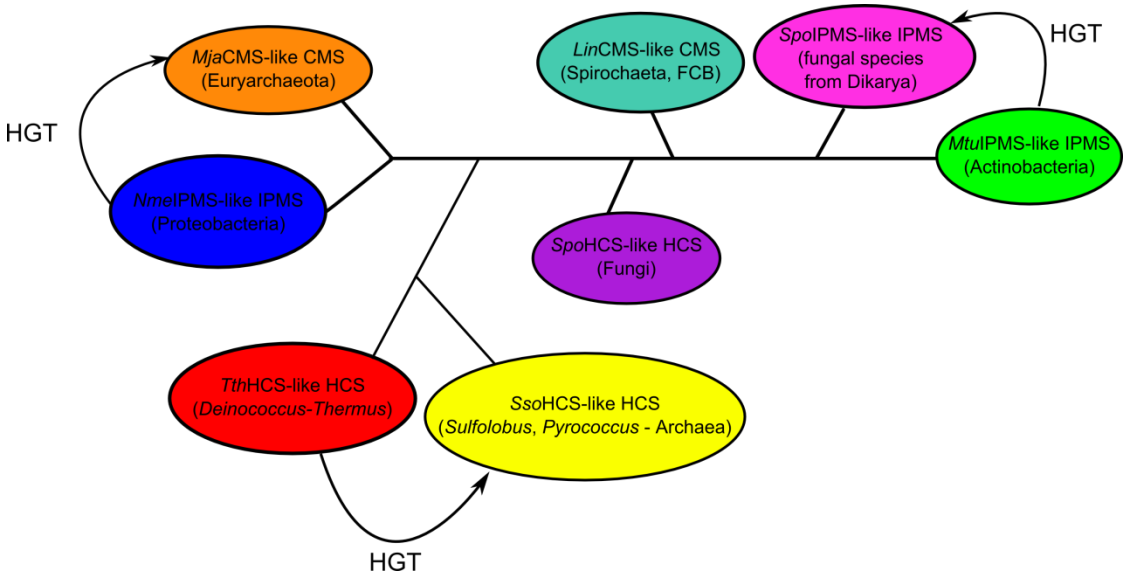


Figure 6.1: A representation of the potential relationships between the different IPMS and IPMS-like enzymes of interest. HGT denotes a potential horizontal gene transfer event.

Based on the pattern of evolution identified by the CLANS analysis in Chapter 3, and work done by Casey et al.[68] and Kumar et al.[69], both horizontal gene transfer, and gene duplication then diversification events are likely to have all played a role in the evolution of the different proteins (Figure 6.1). The *Lin*CMS-like CMSs are evolutionarily distinct from the *Mja*CMS-like CMSs suggesting that this enzymatic activity has risen at least twice. The *Mja*CMS-like CMS are more closely related to the *Nme*IPMS-like IPMS, which are primarily found in Proteobacteria, than they are to the *Lin*CMS-like CMS even though they do not catalyse the same reaction. Taxonomically, the species containing *Mja*CMS-like CMSs are eukaryotes, primarily Euryarchaetoa, but as the proteins show most similarity to IPMSs from Proteobacteria, it suggests there has been a relatively

recent gene transfer event from Proteobacteria that led to the diversification of the *Mja*CMS-like CMSs in those species. Interestingly, there is also evidence of a HCS in methanogenic Archaea, although it appears to be involved in the biosynthesis of coenzyme-B and biotin as opposed to the biosynthesis of lysine.[98]

The taxonomy of the *Lin*CMS-like CMSs also suggests horizontal gene transfer events. The majority of the species that contain a *Lin*CMS-like CMSs are bacteria from the superphylum FCB group, containing Chlorobi, Bacteroidetes, and Fibrobacteria, and bacteria from the Proteobacteria phylum. The phylum Spirochaetes that contains the Leptospira genus is quite distinct from these other phyla, suggesting that there has been a gene transfer event to the *Leptospira* genus as other genera in the Spirochaetes phylum do not contain a CMS sequence and many do not contain an IPMS sequence.

The bacterial HCS, such as that observed in *Thermus thermophilus*, are more closely related to the *Nme*IPMS-like IPMSs than they are to the fungal HCSs, even though there is a significant difference in structure between IPMS and HCS. Nishida et al.[179] analysed phylogenetic relationships inside the *Deinococcus-Thermus* phylum, in which lysine is synthesised primarily through a AAA pathway, and suggested that the entire lysine biosynthetic pathway in this phylum had been transferred to the ancestor of the phylum. Although numerous other phyla have annotated homocitrate synthases in their genomes, they do not have the rest of the aminoadipate pathway, and do have functioning DAP pathways, suggesting that lysine is not synthesised in the same way as it is in the *Deinococcus-Thermus* phylum, where an entire AAA pathway is present.[95]

Figure 6.2: The AAA pathway from fungi, *Deinococcus-Thermus*, and Archaea. The coloured circles represent enzymatic steps in the pathway, with the same colour representing a homologous enzyme in the different organisms. The broken colour in steps 1-3 of the Archaea pathway represents enzyme functionality that has not been identified in all species that have the second half of the pathway in their genomes.

Parts of an AAA pathway is also present in some Archaea, such as *Sulfolobus* and *Pyrococcus*, suggesting another potential gene transfer event (Figure 6.2). The second half of the AAA pathway in bacteria and Archaea is different from that of fungi, as discussed previously, suggesting that the gene transfer event occurred between *Deinococcus-Thermus* and the Archaea that maintain a complete AAA pathway as opposed to a gene transfer event from fungi that utilise different enzymes to convert aminoadipate to lysine. It has also been shown that the small subunit from the homoaconitase from *Pyrococcus horikoshii* can form a functional heterodimer with the large subunit in *Thermus thermophilus* and complement the homoaconitase small subunit knockout in *Thermus thermophilus*, providing further evidence that the AAA pathways from *Deinococcus-Thermus* and Archaea have similar origins.[180] Nishida et al.[85] also argues that there may have been a horizontal

gene transfer event involving the AAA pathway between *Thermus* species and *Pyrococcus horikoshii* in particular, although the phylogenetic tree suggested that this event may have been close to the divergence of Archaea and Bacteria.

The *Mtu*IPMS-like IPMSs are also distinct from the *Nme*IPMS-like IPMSs, and the other proteins. The vast majority of the species containing *Mtu*IPMS-like IPMSs are within the Actinobacteria phylum, while the species containing *Nme*IPMS-like IPMSs are found in other phyla such as Proteobacteria. This may be one of the earliest divisions from the ancestral IPMS. The ancestral protein was likely promiscuous, as this promiscuity is seen in extant enzymes where alternative substrates can be accommodated by the active site of the IPMS. Additionally, the ancestral protein was likely involved in leucine biosynthesis as no alternative pathway for leucine biosynthesis in prokaryotes has been discovered yet there are alternative pathways for both isoleucine and lysine biosynthesis that do not use proteins with the IPMS scaffold. It is also likely that the ancestral IPMS was allosterically regulated, as it appears unlikely that the same unique fold seen in the regulatory domain of both *Mtu*IPMS-like and *Nme*IPMS-like IPMSs arose twice. It has been suggested that Actinobacteria diverged from other bacterial phyla so long ago that the last common ancestor of Actinobacteria and other bacterial phyla can no longer be identified.[181] This substantial difference in sequence between the *Mtu*IPMS-like and *Nme*IPMS-like IPMSs does suggest that divergence of these two groups was comparatively ancient. Therefore, care must be taken when comparing IPMS proteins from the two groups as, although they catalyse the same reaction, they appear to have diverged at the earliest point.

Interestingly, a phylogenetic study was done to investigate the origin of IPMS in fungal species, and this showed that IPMS from fungi such as *Schizosaccharyomyces pombe* and *Saccharomyces cerevisiae*, i.e. fungi from the Dikarya subkingdom of fungi, appear to be more closely related to those IPMS present in Actinobacteria than IPMS from non-Dikarya fungi that appear to share a common origin with IPMS from plants and photosynthetic bacteria.[90] It is not clear whether the origin of the fungal HCS was a different gene transfer event and thus has a different origin to that of the IPMS in the same genome or was the result of gene duplication and divergence from the IPMS pathway in the relevant species.

6.3 Allostery in *Nme*IPMS in the absence of a conformational change

The second question that this work focussed on was about how the allosteric signal was transferred from the regulatory domain to the active site in the apparent absence of a conformational change.

Along with utilising such techniques to explore the broader network that facilitates catalysis, a coevolution analysis was used to investigate the potential for an allosteric network in *Nme*IPMS-like IPMS. Statistical coupling analysis suggested that there was a network of residues from the allosteric binding site to the catalytic barrel. There were significant differences in this network to a network of residues that have been identified as involved in allosteric regulation in *Mtu*IPMS. One residue, Tyr410 (*Mtu*IPMS numbering), was mutated to Phe and this mutant was insensitive to leucine.[63] This residue is highly conserved in all of these proteins. However, the same mutation in *Nme*IPMS had very aberrant kinetics but showed sensitivity to leucine. This difference in the effect of a comparable mutation in *Mtu*IPMS and N*me*IPMS suggests that there is a phylogenetic basis to the network of residues that confer allostery. Based on phylogenetic and taxonomic information, it appears that the Actinobacteria, containing *Mtu*IPMS-like IPMS, and Proteobacteria, containing *Nme*IPMS-like IPMS, diverged early in the evolution of bacteria, and suggests that the two populations have developed different networks to facilitate allosteric regulation using the same structural scaffold.

6.4 The preservation of catalysis does not mean the preservation of allostery

Alanine mutations in *Nme*IPMS of residues in the potential allosteric network have been made. The Arg470Ala and Arg32Ala mutants were insensitive to inhibition by leucine, although ITC showed that leucine was still bound by these proteins. Additionally, another mutant, Glu298Ala, was made, based on that residue's presence in this network, which did not have substantial changes in kinetic parameters compared to the wild type protein yet showed a significant decrease in sensitivity to inhibition by leucine. This suggests that the network of residues determined by SCA may indeed be involved in allosteric signal transmission. Fusion proteins were constructed to determine whether allostery could be transferred by fusion of a regulatory domain to the catalytic unit of *Nme*IPMS. However, although these fusions were catalytically active, they did not show regulation by their respective amino acids. As the regulatory domains were obtained from proteins considerably evolutionarily distinct from *Nme*IPMS, it seems likely that there is a substantial difference in the network of residues that confer allosteric regulation between the two proteins, and therefore allosteric regulation was not able to be transferred.

To explore these fusions further, different fusions could be made such as between *Mja*CMS and *Nme*IPMS, as these proteins share a much closer evolutionary relationship than between *Lin*CMS and *Nme*IPMS, so the network may be maintained between the two proteins. Alternatively, or additionally, mutations could be made in the current fusions to attempt to re-establish allosteric

regulation once it had been determined by ITC whether the allosteric ligand is binding the fusion protein. To further explore the networks of residues identified in Chapter 3 as important for the maintenance of catalysis in the presence and absence of a regulatory domain, mutations could be made in *Nme*IPMS, the truncated form of *Nme*IPMS, or a modern IPMS that lacks a regulatory domain, to determine how changes to this network affect the catalytic activity of the protein, and the dynamics of the subdomains.

6.5 Other avenues in the study of these proteins

There are several additional ways by which the dynamics of *Nme*IPMS in particular could be explored more fully. These include using techniques such as single molecule FRET that would allow for direct detection of the dynamics of this protein. The dynamics could also be explored through a variety of NMR techniques such as H/D exchange as performed with *Mtu*IPMS[78], to determine whether the pathway identified by SCA is identifiable by this method. It could also be of considerable interest to perform a SCA on IPMS from Actinobacteria (i.e. the *Mtu*IPMS-like IPMSs) and assess whether any network suggested by that technique is similar to the network identified by Frantom et al.[78] using H/D exchange.

Another technique that allow for further exploration of the evolution of these proteins is ancestral protein reconstruction. This technique takes modern sequences and, through the construction of a multiple sequence alignment, the construction of a phylogenetic tree, and reconstruction of ancestral sequences using algorithms. The resulting sequences can then be made into synthetic genes and the proteins expressed, purified, and analysed as modern proteins are. This technique could be the ultimate exploration into the evolution of these proteins and their dynamics. The putative *leuA* from *Pyrococcus horikoshii* has also been suggested to code for a bifunctional IPMS and HCS, and also lacks a regulatory domain.[95] The characterisation of this protein may be a unique way to compare functionality with reconstructed ancestral proteins or with more specialised extant enzymes.

6.6 Conclusion

The evolution of protein dynamics is of interest currently as the change in dynamics in response to ligand binding is a comparatively unexplored way by which antibiotic resistance can develop. There are major challenges to investigating dynamics in extant proteins, and new technology may

enable further, deeper, study in the future. This project utilised several techniques to investigate how these very mobile proteins maintained the trade-off between flexibility and stability to confer catalysis, and allostery may alter the dynamics of the proteins in the absence of a conformational change. It is clear that the networks of residues involved in maintaining catalysis and allostery are intertwined and untangling these networks, and how they have evolved, is a considerable task.

Materials and methods

Water

All buffers and solutions were made up with water purified with a Millipore Milli-Q system prior to use.

pH determination

pH of solutions and buffers was determined using a Mettler Toledo™ S220 SevenCompact™ pH/Ion meter. The pH was altered as required using NaOH or HCl solutions.

Structural images of proteins

Structural representations of proteins were generated using the PyMOL Molecular Graphics System (version 1.8.2.1, Schrödinger, LLC)[182].

Media

All *E. coli* cultures were grown in lysogeny broth (LB) at 20 gL^{-1} made up to volume with Milli-Q water and sterilised by autoclave prior to use. Antibiotics were added as required once the media had cooled and prior to use. Agar plates were made using LB media and agar (Miller's, 37 gL^{-1}) made up with Milli-Q and sterilised by autoclave. Following sterilisation, appropriate antibiotics were added once the media had cooled and prior to the plates being poured. SOC media used for transformation was made up of 2% (w/v) tryptone, 0.5% (w/v) yeast extract, 10 mM NaCl, 2.5 mM KCl and 10 mM $MgCl_2$ dissolved in Milli-Q and sterilised by autoclave. 20 mM of filter-sterilised glucose was added once the media had cooled. SOC was stored at -20 °C in aliquots until required.

Cloning

The construct of *Nme*IPMS described by Huisman[64] was used as the template for the construction of site-directed mutants and also for the construction of fusion proteins. The construction of *Nme*IPMS mutants Arg470Ala and Arg32Ala were made as described by Davies[132].

Polymerase chain reaction

PCR for cloning purposes was performed using Phusion® polymerase under standard reaction conditions unless otherwise specified. Colony PCR was performed using *Taq* polymerase (Roche) under standard reaction conditions. PCR was performed in a Veriti® 96-well Thermal Cycler (Applied Biosystems) or in an iCycler (BioRad).

Lyophilised primers for PCR or site-directed mutagenesis were purchased from Invitrogen. The primers were dissolved to the desired concentration in sterile Milli-Q or TE buffer (10 mM Tris, 0.1 mM EDTA, pH 8.0).

InFusion cloning

InFusion cloning was used for cloning genes into vectors in this project. This procedure utilises a proprietary enzyme mixture to clone inserts into vectors with directionality by fusing a 15 bp overlap on the ends of the linear insert to the corresponding sequence in a linearised vector. The vectors used were pET21a or pET28a and were linearised by digestion with restriction enzymes and the linear vector was gel purified using the Nucleospin® Gel and PCR Clean-up kit (Clontech). The ligation reaction was performed under conditions described in the InFusion HD Cloning Kit User Manual. The ligation mixture was then transformed as described below.

Glycerol stock

Glycerol stocks were made using cultures that had been grown overnight at 37°C. A stock solution of 50% glycerol and 50% Milli-Q water was sterilised by autoclave prior to use. 500 µl of overnight culture and 500 µl of the stock glycerol solution were mixed in a sterile Eppendorf tube and flash frozen. The glycerol stocks were stored at -80°C.

Agarose gel electrophoresis

Agarose gels were prepared by mixing 1% (w/v) powdered agarose (Seakem) with TAE buffer (50 mM Tris, 20 mM acetic acid, and 1 mM EDTA), followed by heating until the agarose powder had dissolved. SYBR® Safe gel stain (Invitrogen) was then added at the standard concentration before the gel was cast. Samples were mixed with gel loading buffer (60 mM Tris-HCl, 60 mM EDTA, 0.2% (w/v) Orange G, 0.05% (w/v) xylene cyanol FF, 60% (v/v) glycerol) prior to being loaded into the gel's wells. Gels were typically run at 100 V for 30 minutes or until the dye front had reached a desired position.

If the PCR product was to be used for further cloning, the E-Gel® Safe Imager™ Transilluminator was used to visualise the bands. The band(s) of interest were excised and purification of the DNA was performed using the Nucleospin® Gel and PCR Clean-up kit (Clontech).

Synthetic genes

The *Spo*HCS gene was obtained as a synthetic gene from GeneArt and was codon-optimised for expression in *E. coli*. PCR was used to amplify the gene in the initial generic GeneArt vector, using the primers specified in Materials Table **0.1**. The gene was subsequently sub-cloned into pET28a using the InFusion® HD cloning kit (Clontech) at the NdeI and XhoI sites in the MCS of pET28a. Primers were designed with a 15 bp overlap to allow for use of this technique. The plasmid map of the *Spo*HCS construct is located in Appendix III.

Genes cloned from genomic DNA

The *Sso*HCS gene (*Sso* leuA-2) was amplified from *Sso* genomic DNA using standard Phusion® PCR techniques. As with the *Spo*HCS gene, primers used for PCR amplification included a 15 bp overlap to allow for use of the InFusion HD cloning kit. The *Sso*HCS gene was cloned into pET28a at the NheI and XhoI sites. The plasmid map for the *Sso*HCS construct is located in Appendix III.

Transformation

Following the ligation reaction using the InFusion HD cloning kit or the site-directed mutagenesis reaction, part of the ligation/mutagenesis mixture was transformed into chemically competent Stellar™ *E. coli* cells. These cells were either provided as commercial chemically competent cells or were made chemically competent using the standard $CaCl_2$ method. The ligation mixture was added to a 100 μL aliquot of cells for 30 minutes on ice, followed by a 45 second heat shock at 42°C. Following heat shock, 500 μl of SOC media was added, and the mixture was shaken at 37° for an hour. The mixture was then plated on solid agar plates with appropriate antibiotics added, and were incubated overnight at 37°C.

Materials Table 0.1: Primers designed and used in this project. The complete protein sequences for the fusion constructs are located in Appendix IV,

Enzyme	Primer 5' to 3'
	_Nme_IPMS point mutations
*Nme*IPMS Glu298Ala Forward	CAATGCCTTTTCGCATGCATCGGGCATCCATCAG
*Nme*IPMS Glu298Ala Reverse	CTGATGGATGCCCGATGCATGCGAAAAGGCATTG
*Nme*IPMS Arg371Ala Forward	GAACTCGCCGACAAAAAAGCCGAAATCTTCGATGAAG
*Nme*IPMS Arg371Ala Reverse	CTTCATCGAAGATTTCGGCTTTTTTGTCGGCGAGTTC
	_Nme_IPMS truncation
*Nme*IPMS K395Term Forward	CATGAATGCCGAGAGCTACTAATTCATCTCCCAAAAAATC
*Nme*IPMS K395Term Reverse	GATTTTTTGGGAGATGAATTAGTAGCTCTCGGCATTCATG
	_Nme_IPMS fusions
pET21a – His6 tag overlap primer for use with InFusion cloning	AAGGAGATATACATACATCATCACCATCACCATG
His6-tag and TEV site for cloning into pET21a	CATCATCACCATCACCATGAAAACCTGTATTTTCAGGGCAGCGGCG CG
*Nme*IPMS Forward	CAGGGCAGCGGCGCGATGACACAGACCAACCGCG
*Spo*HCS Forward	CAGGGCAGCGGCGCGATGTCTGTGTCCGAAGCTAATG
*Spo*HCS_G351_Reverse	CAAGCTCAAACGAGAGCCAACATGAACATAAC
*Nme*IPMS_L330_Forward	GGCTCTCGTTTGAGCTTGGGCAAATTGTCCG
*Nme*IPMS_XhoI_Reverse	GGTGGTGGTGCTCGATCAAATCGTACCGCTGCC
*Spo*HCS_T413_Reverse	GATGAATTTGGTGATTCTATCAGCATCACTC
*Nme*IPMS_K395_Forward	GAATCACCAAATTCATCTCCCAAAAAATCAGC
*Nme*IPMS_Reg_Forward	CAGGGCAGCGGCGCGAAATTCATCTCCCAAAAAATCAGC
*Nme*IPMS-*Lin*CMS_Fusion_Reverse	CAGCACTTTTTCACCGCTGCCCATTTCGTCGGATACC
*Lin*CMS_XhoI_Reverse	GGTGGTGGTGCTCGATTAGATTTGCCATGGTTGTAG
*Sso*HCS_NheI_Forward	CAGCCATATGGCTAGCGAAAACCTGTATTTTCAGGGCAGCGGCGC GATGATAAAAGTAGGTATTTTAGATTCGAC
*Sso*HCS_XhoI_Reverse	GGTGGTGGTGCTCGAGTCATTACATTTTCTTAAGGACTAATGATGT
*Nme*IPMS_*Sso*HCS_Fusion_Reverse	CTGATGTGGTGTAGCTCTCGGCATTCATGCT
*Nme*IPMS_*Sso*HCS_Fusion_Forward	CCGAGAGCTACACCACATCAGTAACTCGTCGTTT
	_Spo_HCS cloning
*Spo*HCS_pET28a_Forward	CAGCCATATGGCTAGCGAAAACCTGTATTTTCAGGGCAGCGGCGC GATGAGCGTTAGCGAAGCAAATGGC
*Spo*HCS_pET28a_Reverse	GGTGGTGGTGCTCGATTATGCGCTTGCTTCTTTGGTAATG
	_Sso_HCS cloning
*Sso*HCS_pET28a_Forward	CAGCCATATGGCTAGCGAAAACCTGTATTTTCAGGGCAGCGGCGC GATGATAAAAGTAGGTATTTTAGATTCGAC
*Sso*HCS_pET28a_Reverse	GGTGGTGGTGCTCGAGTCATTACATTTTCTTAAGGACTAATGATGT

Colony PCR

Colonies were screened to assess whether the insert had been ligated into the vector using a gene specific primer and a vector specific primer in a standard *Taq* polymerase PCR reaction. Part of a colony from an agar plate was used as the template for the reaction. Gel electrophoresis, using agarose gel stained with SYBR® Safe, was used to determine if the insert had been ligated into the vector correctly.

Plasmid extraction and sequencing

A single colony, or part of a colony if said colony had been determined to contain a plasmid with the insert correctly inserted into the vector, was selected from an agar plate and used to inoculate a 5 mL culture of liquid LB media. The culture was grown overnight at 37°C with shaking. A glycerol stock was then made from the overnight culture and was stored at -80°C. From the remaining culture, the plasmid was extracted using the ZR Plasmid Miniprep kit (Zymo Research) or the High Pure Plasmid Isolation Kit (Roche). Plasmid concentrations were determined by the absorbance at 260 nm using a NanoDrop® ND-1000 spectrophotometer. Plasmids were sequenced by Macrogen (Korea), the Canterbury Sequencing Facility (University of Canterbury), or the Massey Genome Service (Massey University).

Protein expression and purification

As the vectors used in this study were all pET-based vectors, BL21*(DE3) *E. coli* chemically competent cells were used for transformation of the purified plasmid after the mutation or insert had been confirmed as correct by sequencing. The pET vectors contain a T7 promotor that controls the expression of the gene of interest, while BL21*(DE3) *E. coli* cells contain a λDE3 lysogen that has the gene for a T7 RNA polymerase under the control of a *lac* promoter. Upon induction by isopropyl β-D-1-thiogalactopyranoside (IPTG), the T7 RNA polymerase is expressed and allows the expression of the gene of interest under the control of the T7 promoter.

Protein expression strain creation

Chemically competent BL21*(DE3) cells were transformed with the plasmid of interested as detailed above. Following isolation of individual colonies on agar plates, an overnight or pre-culture was inoculated and grown overnight for creation of a glycerol stock or for protein expression.

Cell cultures

Pre-culture of 20 – 100 mL of LB media with appropriate antibiotics were inoculated with the expression cell strain of interest and was grown overnight at 37°C with shaking. A pre-culture was then added to a large culture (1 L of LB media with appropriate antibiotics in a 2 L baffled flask). Large cultures were then grown at 37°C with shaking until the optical density OD at 600 nm (OD_{600}) had reached 0.4 – 0.8 AU. IPTG was then added to the cultures to a final concentration of 0.5 mM. Cultures were then grown overnight at 23°C with shaking, or for 4 hours at 37°C with shaking.

Large cultures were harvested by centrifugation using 1 L bottles in a Fiberlite™ F9-6x1000 LEX fixed-angle rotor (ThermoFisher Scientific). The cells were pelleted at 14000g for 30 mins at 4°C. The cell pellets were subsequently stored in sterile 50 mL tubes at -80°C until required.

Cell lysis

Cell lysis was performed using a Omni-Ruptor 4000 Ultrasonic Homogeniser sonicator with 4–6 repeats of 5 minutes at 70% power with a pulse of 50%. The cell pellet was re-suspended in 20 – 40 mL of cold lysis buffer and lysis was performed on ice. Benzonase® was added to the lysate prior to centrifugation at 40000 g for 45 minutes. The buffers used for purification of specific proteins are detailed in Materials Materials Table **0.2**, Materials Materials Table **0.3**, and Materials Materials Table **0.4**.

Fast protein liquid chromatography (FPLC)

FPLC was performed at 4°C using a Bio-Rad Biologic Protein Chromatography system or an ÄKTApurifier™ (GE Healthcare). Buffers were filtered through a 0.2 μm filter and cooled to 4°C prior to use. Protein lysate was also filtered through a 0.2 μm filter prior to loading into a Superloop™(GE Healthcare).

Immobilised metal affinity chromatography was performed as the first purification step. The 5 mL HisTrap column was equilibrated with equilibration buffer after the storage solution of 20% ethanol (v/v) had been removed. The sample was then loaded onto the column, the sample was washed with equilibration buffer to removed non-specific binding, and a gradient from 20 mM imidazole to 500 mM imidazole was used to elute the protein of interest from the column as 2 mL fractions. The purification was followed by tracking absorbance at 280 nm, and fractions that

showed absorbance during elution were analysed by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE).

Following IMAC, the protein was concentrated by centrifugation using a Vivaspin™ 20 mL 10 kDa cut-off concentrator. Once the protein had reached the appropriate volume or concentration, the protein was loaded onto a HiLoad™ 26/60 Superdex™ 200 prep grade size exclusion column that had been equilibrated into the appropriate SEC buffer. The protein was then eluted from the column with a column volume of SEC buffer and was collected as 2 mL aliquots. As with the IMAC, the progress of the purification was followed by tracking the absorbance at 280 nm, and fractions that showed substantial absorbance at 280 nm were analysed by SDS-PAGE.

The purified protein was then stored as 50 – 500 μL aliquots at a concentration of 1 – 10 $mg.mL^{-1}$ at -80°C after being flash frozen in liquid nitrogen.

Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE)

SDS-PAGE was performed using a precast Bolt® 10% Bis-Tris Plus gel (Invitrogen). All gels were run in MOPS and were run at 200 V for 30 minutes. Protein samples were prepared as described in the product manual. Gels were stained using a heated solution of 10% (v/v) glacial acetic acid, and 40% (v/v) methanol and 0.1% (w/v) Coomassie Brilliant Blue R-250 until the gel was sufficiently stained. Gels were then de-stained using a heated solution of 10% (v/v) glacial acetic acid and 40% (v/v) methanol until bands of interest were appropriately visible.

Determination of protein concentration

The approximate extinction co-efficient of the protein of interest was determined using the ProtParam tool on the ExPASy server.[183] Protein concentration was approximated using these values and the absorbance at 280 nm measured using a Nanodrop® ND-1000 spectrophotometer.

Enzyme kinetic assays

The assay used for determination of enzyme kinetics was the previously described assay using 4'-4'-dithiopyridine (DTP) to detect the formation of CoA at 324 nm ($\varepsilon = 1.98 \times 10^4\ M^{-1}cm^{-1}$).[1, 74, 79] The assays were performed in stoppered quartz cuvettes with a path length of 1 cm, and the data was collected using a Varian Cary 100 UV-visible spectrophotometer at 25°C for all proteins aside from *Sso*HCS, where the temperature used was 60°C. Measurements were made in duplicate.

The buffer used for kinetic experiments was 50 mM HEPES, pH 7.5, 20 mM KCl and 20 mM $MgCl_2$. Stock solutions of substrates at 50 mM were used as required to achieve the concentration required. The stock solution of AcCoA was stored at -80°C. Typically, the buffer, AcCoA, DTP, and enzyme at the concentration required were added to the cuvette, and once the residual CoA had reacted with the DTP, the reaction was initiated with the addition of the other substrate (KIV/KG).

Inhibition data was obtained using the standard enzyme assay with different concentrations of inhibitor. The concentration of substrates was held at saturating concentrations for the enzyme of interest.

The concentration of substrates was also determined using this assay. A limiting amount of one substrate was added to the assay while the other substrate was held in excess. As with the enzyme assays, the buffer, DTP, AcCoA, and enzyme were added to the cuvette first and the reaction was initiated by the addition of the other substrate once there was no change in absorbance caused by excess CoA due to hydrolysis of AcCoA to CoA. The change in absorbance was measured, and this was used to calculate the concentration of the limiting substrate using the Beer-Lambert Law. The extinction coefficient of the reaction product is $1.98 \times 10^4\ M^{-1}cm^{-1}$.

Differential scanning fluorimetry (DSF)

DSF was performed using a iCycler iQ5 Multicolor Real-Time PCR Detection System (Bio-Rad) to determine the melting temperatures (T_m) of proteins of interest. Protein samples ($0.1\ mg.mL^{-1}$) were mixed with SEC buffer and SYPRO® Orange protein gel stain in the presence of substrate (250 μM KIV or KG) or inhibitor (1 mM L-Leu, L-Ile, or L-Lys). Controls that included buffer instead of the protein sample were also prepared. The temperature increased at a rate of 1°C/minute from 20° to 98°C. Fluorescence was measured at 0.5°C intervals. The controls were subtracted from the protein samples to remove background fluorescence. The T_m was derived from the inflection point of the sigmoidal graph on a plot of the fluorescence intensity as a function of temperature.

Analytical size exclusion chromatography (analytical SEC)

Analytical SEC was used to assess the oligomeric state of *Nme*IPMS K395Term. As detailed in Chapter 4.1.4, several different buffers were used as the oligomeric state of *Nme*IPMS K395Term was affected by the concentration of salt in the buffer. A Superdex™ 200 10/300 GL column (GE Healthcare) was used to perform analytical SEC. The column was equilibrated

in the buffer required, and protein samples or protein standards at 1 mg/mL^{-1} unless stated otherwise were injected onto the column in a volume of 500 μL. Blue dextran was used to calculate the void volume of the column. Bovine serum albumin (BSA, 66 kDa), apoferritin (443 kDa), cytochrome c (12.4 kDa), β-amylase (200 kDa), and coalbumin (75 kDa) were used as known molecular weight standards. The elution volume for the protein of interest and the molecular weight standards were all recorded, and a linear graph of the log[protein standard mass (Da)] against the elution volume of the standard was obtained. From this, the oligomeric state of the protein of interest could be approximated.

Small-angle X-ray scattering (SAXS)

Small-angle X-ray scattering data was obtained for wild-type *Nme*IPMS and *Nme*IPMS K395Term. Measurements were obtained at the Australian Synchrotron SAXS/WAXS beamline equipped with a Pilatus detector. The wavelength of the X-rays was 1.0332 Å and the sample-detector distance was 1.6 m.

Scattering data for the proteins of interest was obtained following elution from a Superdex® 200 Increase 5/150 size-exclusion column that had been equilibrated in the *Nme*IPMS SEC buffer. This buffer also had the addition of 250 μM KIV or 1 mM L-Leu or 200 μM L-Leu as appropriate for the experiment in particular. All buffers also had 5% glycerol added to limit radiation damage.

Raw data was processed, and the background was subtracted, using Scatterbrain (Australian Synchrotron). The scattering from peaks that showed substantial absorbance at 280 nm were summed and averaged. Plots of scattering intensity (*I*) versus *s* and Guinier plots were generated using Primus.[184] The plots were assessed for increasing intensity at low *s* that is indicative of aggregation. Indirect Fourier transform was performed using GNOM to generate the *P(r)* function.[185] Crysol was used to generate theoretical scattering curves for the *Nme*IPMS homology model and other models produced by molecular dynamics simulations.[157]

X-ray crystallography trials

Attempts were made to crystallise *Nme*IPMS K395Term in the presence and absence of KIV. A Mosquito® Crystal robot (TTP Labtech) was used to screen for potential conditions in which the protein would crystallise using the sitting-drop vapour diffusion technique. The screen conditions tested were PACT premier™ HT-96, Clear Strategy™ I HT-96 and Clear Strategy™ II HT-96, and JCSG-plus™ HT-96 (Molecular Dimensions). Protein concentrations of 5 – 30

mg.ml^{-1} were tested, and the ligand, if added, was added to the protein prior to the protein being mixed with the condition. 400 nL of protein sample was mixed with 400 nL of the condition of interest, and 40 nL of this was used as the reservoir.

Isothermal titration calorimetry (ITC)

ITC was performed using a Nano ITC Low Volume (TA Instruments). Experiments were performed at 25°C unless otherwise specified. Purified protein samples of 400 μL at 10 – 15 mg.ml^{-1} were de-gassed and injected into the cell after the cell had been washed with de-gassed buffer from the same batch that the protein was purified in. The ligand solution was made up in the same buffer. The syringe was washed with de-gassed buffer and then de-gassed ligand solution before the ligand sample was loaded into the syringe. 2 μL of ligand was injected into the sample cell every 200 s. ITC parameters were determined using the NanoAnalyze software (TA Instruments).

Multiple sequence alignments

Sequence populations were obtained from KEGG[172-174], Pfam[114, 115], the NCBI Protein database, and PSI-BLAST[101]. CD-HIT or the CD-HIT online webserver was used to filter the sequence populations to remove redundancy.[125, 186] Multiple sequence alignments were performed using the MAFFT online server.[187] The choice in algorithm is detailed in Chapter 2.2.3. MSAs were visualised using Jalview and were manually edited to remove aberrant sequences from the alignments.[188]

Cluster Analysis of Sequences (CLANS)

CLANS was performed on different sequence populations.[126] As described in Chapter 2.2.2, CLANS uses an all versus all BLAST search to calculate pairwise attraction values, and these values are then used to create a force-directed graph to observe clusters of sequences depending on their relationship to each other. The input is sequences in FASTA format and clustering can be performed using a variety of methods. Network-based clustering with a minimum of 10 sequences per cluster was typically used to determine the clusters in the sequence population. The different clusters were then analysed to investigate which sequences were contained within the clusters to assign them to a known group within the sequence populations.

Statistical coupling analysis (SCA)

The SCA was performed in MATLAB using sca5.m. Trimming of the MSA and further steps were performed as described in Chapters 2 and 3.

Mutual information (MIp)

Covariance analysis using MIp was performed as described in Chapter 3, using Linux Ubuntu 16.04 and The MIp Toolset.[144] A coevolution network file was produced and visualised using Graphviz.[189] The dist_pdb programme within the MIp toolset was used to obtain the distances between atoms in the *Nme*IPMS homology model PDB file. BioPython and code adapted from https://warwick.ac.uk/fac/sci/moac/people/students/peter_cock/python/protein_contact_map/ was used to generate contact maps using these distances.[190] The graphical images were produced using MATLAB.

Materials Table 0.2: Buffers used for the purification of *Nme*IPMS and variants

Buffers used for the purification of *Nme*IPMS and variants		
Buffer	*Other components*	*Uses*
50 mM potassium phosphate, pH 8.0	300 mM KCl, 20 mM imidazole	Lysis, and equilibration for HisTrap
50 mM potassium phosphate, pH 8.0	300 mM KCl, 500 mM imidazole	HisTrap elution
50 mM HEPES, pH 7.5	20 mM KCl, 20 mM $MgCl_2$ 5% (v/v) glycerol (SAXS) 0 – 30% (v/v) glycerol (viscosity-dependent kinetics)	HiLoad™ 26/60 Superdex™ 200 prep grade column size exclusion chromatography buffer. This buffer was also used for SAXS, kinetics, and ITC for *Nme*IPMS and mutants as detailed.
10 mM Tris, pH 8.0	300 mM KCl	Superdex™ 200 10/300 GL column size exclusion buffer Used for analytical SEC.

Materials Table 0.3: Buffers used for the purification of *Spo*HCS and variants

Buffers used for the purification of *Spo*HCS and variants		
Buffer	*Other components*	*Uses*
50 mM sodium phosphate, pH 7.0	500 mM NaCl, 20 mM imidazole	Lysis, and equilibration for HisTrap
50 mM sodium phosphate, pH 7.0	500 mM NaCl, 500 mM imidazole	HisTrap elution

Materials Table 0.4: Buffers used for the purification of *Sso*HCS

Buffers used for the purification of *Sso*HCS		
Buffer	*Other components*	*Uses*
50 mM Tris, pH 8.5	100 mM NaCl, 20 mM imidazole	Lysis, and equilibration for HisTrap
50 mM Tris, pH 8.5	100 mM NaCl, 500 mM imidazole	HisTrap elution
50 mM HEPES, pH 7.5	20 mM KCl, 20 mM $MgCl_2$	Kinetics

Appendix I

Residues (*Nme*IPMS numbering) within the independent components identified in Chapter 3

Independent component 1 (IC1)
10
49
56
57
64
93
97
101
104
161
169
170
176
209
226
228
229
253
296
303
315
316
336
338
339
360
363
459
480
495
498

Independent component 2 (IC2)
10
21
47
48
76
77
82
86
99
104
112
120
139
141
145
147
153
163
164
165
174
194
202
205
210
212
224
244
245
248
279
288
294
297
298
299
307
312
314
321
329
333
364
371

Independent component 2 (IC2) continued
375
380
396
431
432
458
460
467
469
491

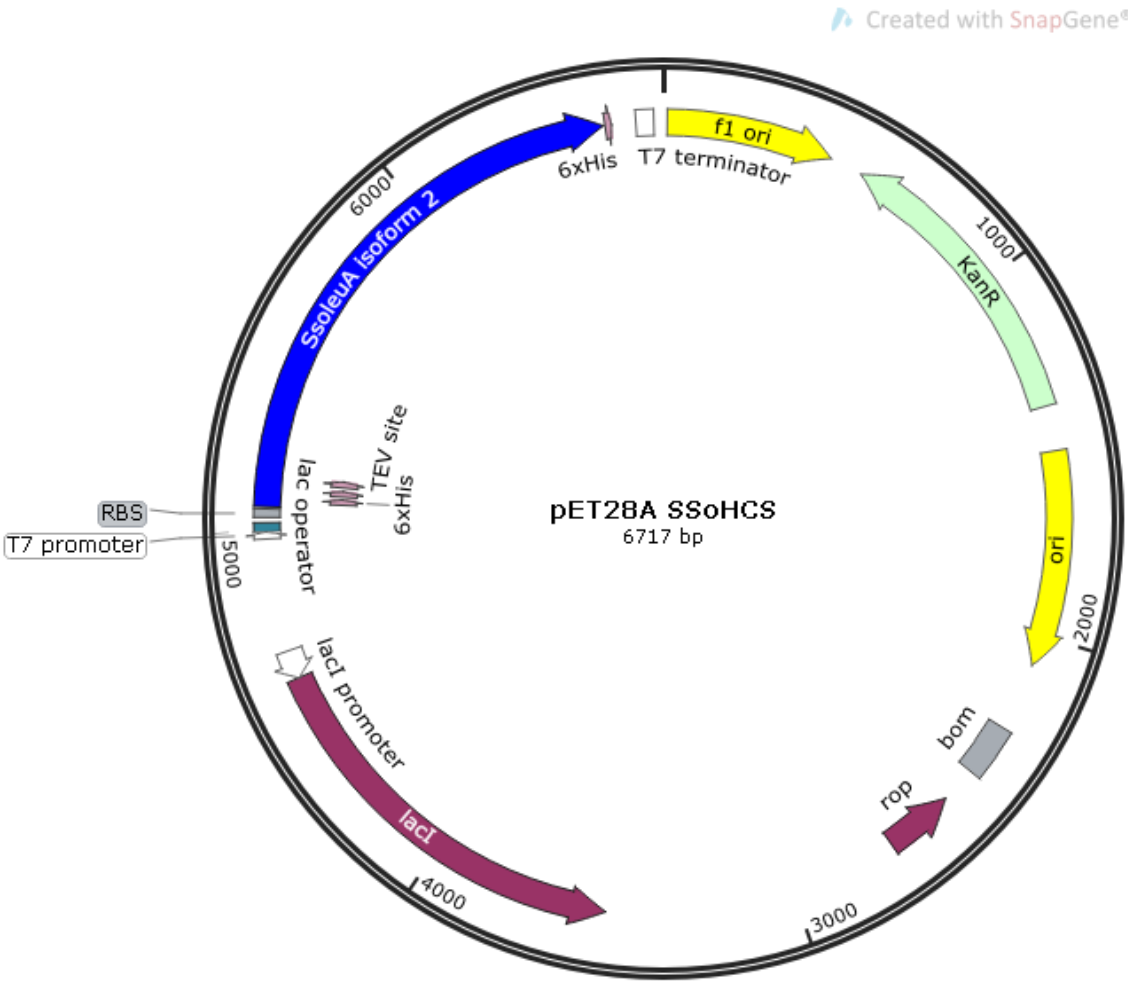
Independent component 3 (IC3)
29
60
73
75
84
130
171
181
194
247
248
285
304
308
334
346
370
404
410
431
433
456
487

Appendix II

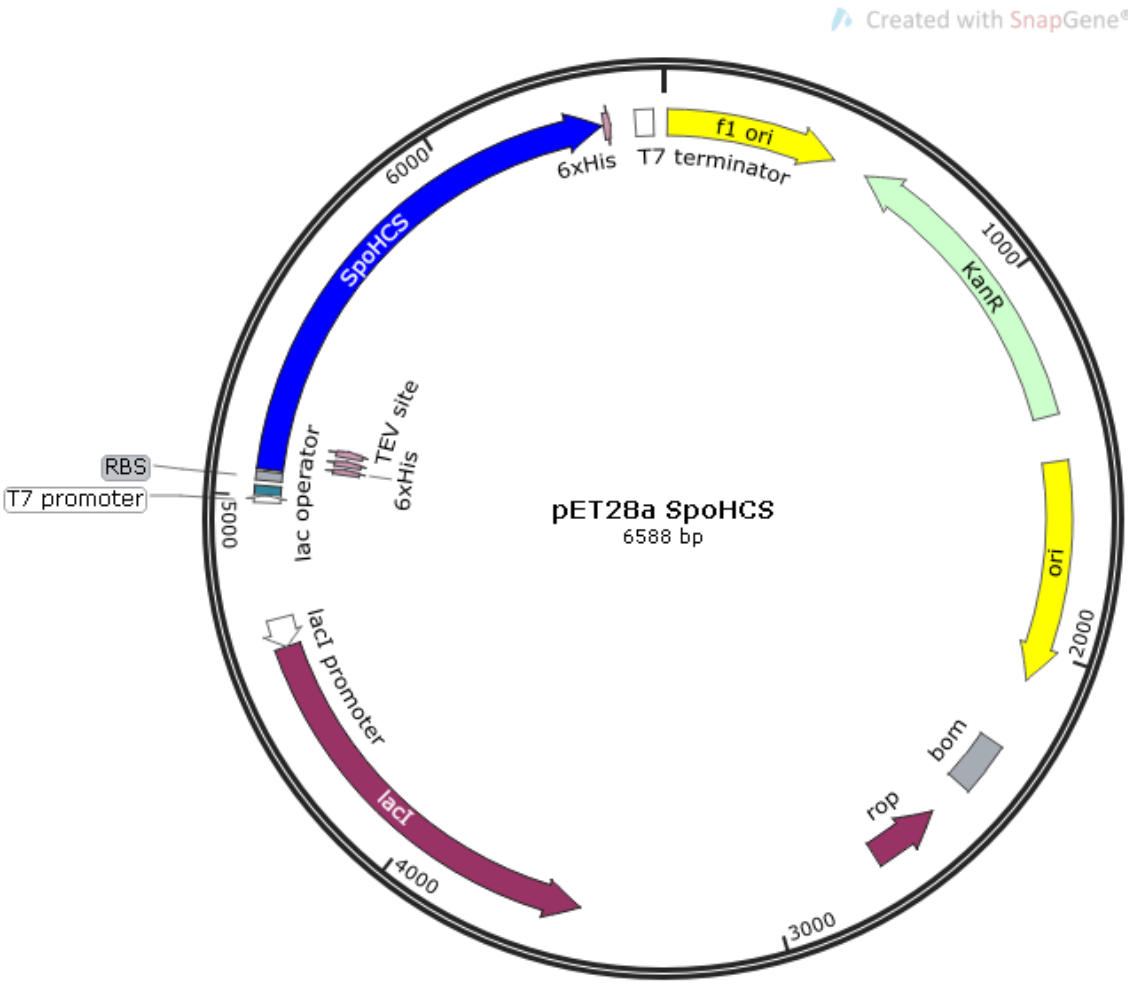
Residues (*Nme*IPMS numbering) within the principal component of the RDA alignment identified in Chapter 3

Principal component residues from the RDA SCA
23
25
29
43
45
52
56
57
68
75
83
94
99
101
108
119
135
146
167
169
180
202
205
231
232
241
278
293
298
314
331
332
333
360
364
371

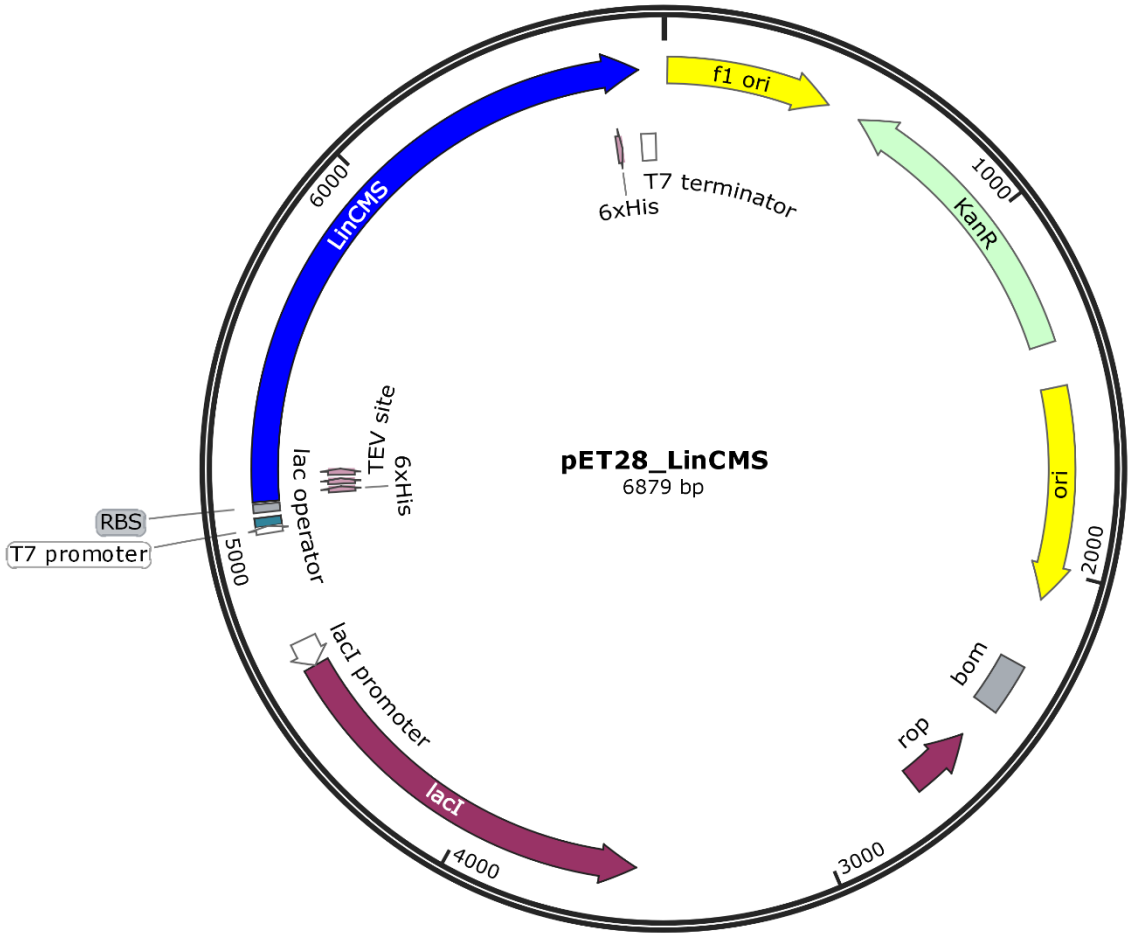
Appendix III



Plasmid map of the *Sso*HCS construct described in Chapter 5.



Plasmid map of the *Sso*HCS construct described in Chapter 5.



Plasmid map of the *Lin*CMS construct described in Chapter 5.

Appendix IV

*Nme*IPMS and *Spo*HCS fusions

The part of the fusion protein that is from *Nme*IPMS is coloured blue, the part that is from *Spo*HCS is coloured in purple.

*Spo*HCS$_{Cat-SI}$ – *Nme*IPMS $_{SII-Reg}$ contains *Spo*HCS from residue 1 to 351 (*Spo*HCS numbering) and *Nme*IPMS from residues 330 to 517 (*Nme*IPMS numbering).

>*Spo*HCS$_{Cat-SI}$–*Nme*IPMS$_{SII-Reg}$
MSVSEANGTETIKPPMNGNPYGPNPSDFLSRVNNFSIIESTLREGEQFANAFFDTEKKIQ
IAKALDNFGVDYIELTSPVASEQSRQDCEAICKLGLKCKILTHIRCHMDDARVAVETGVD
GVDVVIGTSQYLRKYSHGKDMTYIIDSATEVINFVKSKGIEVRFSSEDSFRSDLVDLLSL
YKAVDKIGVNRVGIADTVGCATPRQVYDLIRTLRGVVSCDIECHFHNDTGMAIANAYCAL
EAGATHIDTSILGIGERNGITPLGALLARMYVTDREYITHKYKLNQLRELENLVADAVEV
QIPFNNYITGMCAFTHKAGIHAKAILANPSTYEILKPEDFGMSRYVHVGSRLSLGKLSGR
NAFKTKLADLGIELESEEALNAAFARFKELADKKREIFDEDLHALVSDEMGSMNAESYKF
ISQKISTETGEEPRADIVFSIKGEEKRASATGSGPVDAIFKAIESVAQSGATLQIYSVNA
VTQGTESQGETSVRLARGNRVVNGQGADTDVLVATAKAYLSALSKLEFSAAKPKAQGSGT
I

*Spo*HCS$_{cat}$ – *Nme*IPMS$_{SDs-Reg}$ contains *Spo*HCS from residue 1 to 413 (*Spo*HCS numbering) and *Nme*IPMS from residues 395 – 517 (*Nme*IPMS numbering).

>*Spo*HCS$_{cat}$–*Nme*IPMS$_{SDs-Reg}$
MSVSEANGTETIKPPMNGNPYGPNPSDFLSRVNNFSIIESTLREGEQFANAFFDTEKKIQ
IAKALDNFGVDYIELTSPVASEQSRQDCEAICKLGLKCKILTHIRCHMDDARVAVETGVD
GVDVVIGTSQYLRKYSHGKDMTYIIDSATEVINFVKSKGIEVRFSSEDSFRSDLVDLLSL
YKAVDKIGVNRVGIADTVGCATPRQVYDLIRTLRGVVSCDIECHFHNDTGMAIANAYCAL
EAGATHIDTSILGIGERNGITPLGALLARMYVTDREYITHKYKLNQLRELENLVADAVEV
QIPFNNYITGMCAFTHKAGIHAKAILANPSTYEILKPEDFGMSRYVHVGSRLTGWNAIKS
RAEQLNLHLTDAQAKELTVRIKKLADVRTLAMDDVDRVLREYHADLSDADRITKFISQKI
STETGEEPRADIVFSIKGEEKRASATGSGPVDAIFKAIESVAQSGATLQIYSVNAVTQGT
ESQGETSVRLARGNRVVNGQGADTDVLVATAKAYLSALSKLEFSAAKPKAQGSGTI

*Nme*IPMS and *Lin*CMS fusion

The part of the fusion protein from *Nme*IPMS, from residue 1 to 388 (*Nme*IPMS numbering) is coloured blue. The part of the fusion protein from *Lin*CMS, from residue 388 to 516 (*Lin*CMS numbering) is coloured green.

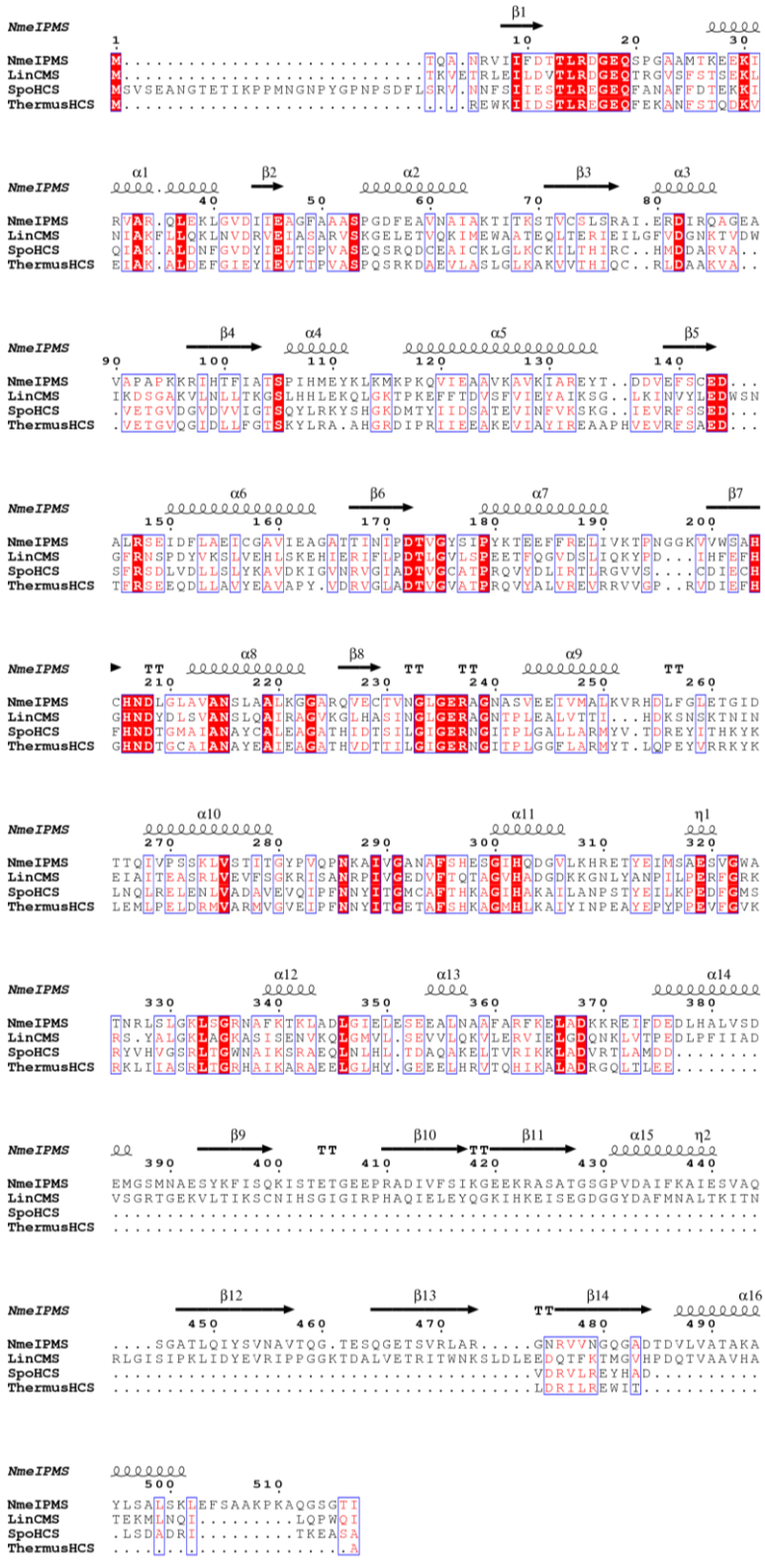
>*Nme*IPMS$_{Cat\text{-}SDs}$ –*Lin*CMS$_{Reg}$
MTQANRVIIFDTTLRDGEQSPGAAMTKEEKIRVARQLEKLGVDIIEAGFAAASPGDFEAVNA
IAKTITKSTVCSLSRAIERDIRQAGEAVAPAPKKRIHTFIATSPIHMEYKLKMKPKQVIEAA
VKAVKIAREYTDDVEFSCEDALRSEIDFLAEICGAVIEAGATTINIPDTVGYSIPYKTEEFF
RELIVKTPNGGKVVWSAHCHNDLGLAVANSLAALKGGARQVECTVNGLGERAGNASVEEIVM
ALKVRHDLFGLETGIDTTQIVPSSKLVSTITGYPVQPNKAIVGANAFSHESGIHQDGVLKHR
ETYEIMSAESVGWATNRLSLGKLSGRNAFKTKLADLGIELESEEALNAAFARFKELADKKRE
IFDEDLHALVSDEMGS**G**EKVLTIKSCNIHSGIGIRPHAQIELEYQGKIHKEISEGDGGYDAF
MNALTKITNRLGISIPKLIDYEVRIPPGGKTDALVETRITWNKSLDLEEDQTFKTMGVHPDQ
TVAAVHATEKMLNQILQPWQI

*Nme*IPMS and *Sso*HCS fusion

The part of the fusion from *Nme*IPMS, from residues 1 to 394, is coloured in blue. The part of the fusion from *Sso*HCS, from residues 399 to 461, is shown in red.

>*Nme*IPMS$_{Cat\text{-}SDs}$–*Sso*HCS$_{Reg}$
MTQANRVIIFDTTLRDGEQSPGAAMTKEEKIRVARQLEKLGVDIIEAGFAAASPGDFEAVNA
IAKTITKSTVCSLSRAIERDIRQAGEAVAPAPKKRIHTFIATSPIHMEYKLKMKPKQVIEAA
VKAVKIAREYTDDVEFSCEDALRSEIDFLAEICGAVIEAGATTINIPDTVGYSIPYKTEEFF
RELIVKTPNGGKVVWSAHCHNDLGLAVANSLAALKGGARQVECTVNGLGERAGNASVEEIVM
ALKVRHDLFGLETGIDTTQIVPSSKLVSTITGYPVQPNKAIVGANAFSHESGIHQDGVLKHR
ETYEIMSAESVGWATNRLSLGKLSGRNAFKTKLADLGIELESEEALNAAFARFKELADKKRE
IFDEDLHALVSDEMGSMNAESYTTSVTRRLSVINGVKEVMEISGDYDILVKVQAKDSNELNQ
IIESIRATKGVRSTLTSLVLKKM

Appendix V



References

1. Huisman, F.H.A., M.F.C. Hunter, S.R.A. Devenish, J.A. Gerrard, and E.J. Parker, *The C-terminal regulatory domain is required for catalysis by Neisseria meningitidis alpha-isopropylmalate synthase.* Biochem. Biophys. Res. Commun., 2010. **393**(1): p. 168-173.
2. Zhang, Z., J. Wu, W. Lin, J. Wang, H. Yan, W. Zhao, J. Ma, J. Ding, P. Zhang, and G.-P. Zhao, *Subdomain II of a-Isopropylmalate Synthase is Essential for Activity: Inferring a Mechanism of Feedback Inhibition.* J. Biol. Chem., 2014.
3. Eisenmesser, E.Z., O. Millet, W. Labeikovsky, and D.M. Korzhnev, *Intrinsic dynamics of an enzyme underlies catalysis.* Nature, 2005. **438**(7064): p. 117-21.
4. Klinman, J.P. and A. Kohen, *Hydrogen Tunneling Links Protein Dynamics to Enzyme Catalysis.* Annu. Rev. Biochem., 2013. **82**(1): p. 471-496.
5. Luk, L.Y.P., J. Javier Ruiz-Pernía, W.M. Dawson, M. Roca, E.J. Loveridge, D.R. Glowacki, J.N. Harvey, A.J. Mulholland, I. Tuñón, V. Moliner, and R.K. Allemann, *Unraveling the role of protein dynamics in dihydrofolate reductase catalysis.* Proc. Natl. Acad. Sci. U. S. A., 2013. **110**(41): p. 16344-16349.
6. Choy, M.S., Y. Li, L.E.S.F. Machado, M.B.A. Kunze, C.R. Connors, X. Wei, K. Lindorff-Larsen, R. Page, and W. Peti, *Conformational Rigidity and Protein Dynamics at Distinct Timescales Regulate PTP1B Activity and Allostery.* Mol. Cell, 2017. **65**(4): p. 644-658.e5.
7. Tobi, D. and I. Bahar, *Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state.* Proc. Natl. Acad. Sci. U. S. A., 2005. **102**(52): p. 18908-18913.
8. Dong, M., S. Husale, and O. Sahin, *Determination of protein structural flexibility by microsecond force spectroscopy.* Nat. Nanotech., 2009. **4**(8): p. 514-7.
9. Davis, I.W., W.B. Arendall, D.C. Richardson, and J.S. Richardson, *The backrub motion: how protein backbone shrugs when a sidechain dances.* Structure, 2006. **14**(2): p. 265-274.
10. Sawaya, M.R. and J. Kraut, *Loop and subdomain movements in the mechanism of Escherichia coli dihydrofolate reductase: crystallographic evidence.* Biochemistry, 1997. **36**(3): p. 586-603.
11. Avlani, V.A., K.J. Gregory, C.J. Morton, M.W. Parker, P.M. Sexton, and A. Christopoulos, *Critical Role for the Second Extracellular Loop in the Binding of Both Orthosteric and Allosteric G Protein-coupled Receptor Ligands.* J. Biol. Chem., 2007. **282**(35): p. 25677-25686.
12. Cross, P.J., R.C.J. Dobson, M.L. Patchett, and E.J. Parker, *Tyrosine Latching of a Regulatory Gate Affords Allosteric Control of Aromatic Amino Acid Biosynthesis.* J. Biol. Chem., 2011. **286**(12): p. 10216-10224.
13. Fushman, D., *Determining protein dynamics from (15)N relaxation data by using DYNAMICS.* Methods Mol. Biol., 2012. **831**: p. 485-511.
14. Mangia, S., N.J. Traaseth, G. Veglia, M. Garwood, and S. Michaeli, *Probing slow protein dynamics by adiabatic R 1ρ and R 2ρ NMR experiments.* J. Am. Chem. Soc., 2010. **132**(29): p. 9979-9981.
15. Fenwick, R.B., D. Oyen, and P.E. Wright, *Multi-probe relaxation dispersion measurements increase sensitivity to protein dynamics.* Phys. Chem. Chem. Phys., 2016. **18**(8): p. 5789-5798.
16. Oyen, D., R.B. Fenwick, P.C. Aoto, R.L. Stanfield, I.A. Wilson, H.J. Dyson, and P.E. Wright, *Defining the Structural Basis for Allosteric Product Release from E. coli Dihydrofolate Reductase Using NMR Relaxation Dispersion.* J. Am. Chem. Soc., 2017. **139**(32): p. 11233-11240.
17. Karplus, M. and J.A. McCammon, *Molecular dynamics simulations of biomolecules.* Nat. Struct. Mol. Biol., 2002. **9**(9): p. 646-652.
18. Zhao, G., J.R. Perilla, E.L. Yufenyuy, X. Meng, B. Chen, J. Ning, J. Ahn, A.M. Gronenborn, K. Schulten, C. Aiken, and P. Zhang, *Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics.* Nature, 2013. **497**(7451): p. 643-646.

19. Estabrook, R.A., J. Luo, M.M. Purdy, V. Sharma, P. Weakliem, T.C. Bruice, and N.O. Reich, *Statistical coevolution analysis and molecular dynamics: Identification of amino acid pairs essential for catalysis.* Proc. Natl. Acad. Sci. U. S. A., 2005. **102**(4): p. 994-999.
20. Nettels, D., A. Hoffmann, and B. Schuler, *Unfolded protein and peptide dynamics investigated with single-molecule FRET and correlation spectroscopy from picoseconds to seconds.* J. Phys. Chem. B, 2008. **112**(19): p. 6137-6146.
21. Mauldin, R.V., M.J. Carroll, and A.L. Lee, *Dynamic dysfunction in dihydrofolate reductase results from antifolate drug binding: modulation of dynamics within a structural state.* Structure, 2009. **17**(3): p. 386-394.
22. Peng, J.W., *Communication Breakdown: Protein Dynamics and Drug Design.* Structure, 2009. **17**(3): p. 319-320.
23. Podust, L.M., T.L. Poulos, and M.R. Waterman, *Crystal structure of cytochrome P450 14a-sterol demethylase (CYP51) from Mycobacterium tuberculosis in complex with azole inhibitors.* Proc. Natl. Acad. Sci. U. S. A., 2001. **98**(6): p. 3068-3073.
24. Rose, R.B., C.S. Craik, and R.M. Stroud, *Domain flexibility in retroviral proteases: structural implications for drug resistant mutations.* Biochemistry, 1998. **37**(8): p. 2607-2621.
25. MacLennan, I.C., *Germinal centers.* Annu. Rev. Immunol., 1994. **12**: p. 117-39.
26. Zimmermann, J., E.L. Oakman, I.F. Thorpe, X. Shi, P. Abbyad, C.L. Brooks, S.G. Boxer, and F.E. Romesberg, *Antibody evolution constrains conformational heterogeneity by tailoring protein dynamics.* Proc. Natl. Acad. Sci. U. S. A., 2006. **103**(37): p. 13722-13727.
27. Adhikary, R., W. Yu, M. Oda, J. Zimmermann, and F.E. Romesberg, *Protein dynamics and the diversity of an antibody response.* J. Biol. Chem., 2012. **287**(32): p. 27139-27147.
28. Boucher, J.I., J.R. Jacobowitz, B.C. Beckett, S. Classen, and D.L. Theobald, *An atomic-resolution view of neofunctionalization in the evolution of apicomplexan lactate dehydrogenases.* eLife, 2014. **3**: p. e02304.
29. Tsai, C.-J., A. del Sol, and R. Nussinov, *Allostery: Absence of a Change in Shape Does Not Imply that Allostery Is Not at Play.* J. Mol. Biol., 2008. **378**(1): p. 1-11.
30. Peracchi, A. and A. Mozzarelli, *Exploring and exploiting allostery: Models, evolution, and drug targeting.* Biochim. Biophys. Acta, 2011. **1814**(8): p. 922-933.
31. Thomas, S. and D.A. Fell, *Design of Metabolic Control for Large Flux Changes.* J. Theor. Biol., 1996. **182**(3): p. 285-298.
32. Goyal, S., J. Yuan, T. Chen, J.D. Rabinowitz, and N.S. Wingreen, *Achieving Optimal Growth through Product Feedback Inhibition in Metabolism.* PLoS Comput. Biol., 2010. **6**(6): p. e1000802.
33. Mathonet, P., H. Barrios, P. Soumillion, and J. Fastrez, *Selection of allosteric β-lactamase mutants featuring an activity regulation by transition metal ions.* Protein Sci., 2006. **15**(10): p. 2335-2343.
34. Ostermeier, M. and S.J. Benkovic, *Evolution of protein function by domain swapping.* Adv. Protein Chem, 2000. **55**: p. 29-77.
35. Cross, P.J., T.M. Allison, R.C.J. Dobson, G.B. Jameson, and E.J. Parker, *Engineering allosteric control to an unregulated enzyme by transfer of a regulatory domain.* Proc. Natl. Acad. Sci. U. S. A., 2013. **110**(6): p. 2111-2116.
36. Ostermeier, M., *Engineering allosteric protein switches by domain insertion.* Protein Eng. Des. Sel., 2005. **18**(8): p. 359-364.
37. Guntas, G. and M. Ostermeier, *Creation of an allosteric enzyme by domain insertion.* J. Mol. Biol., 2004. **336**(1): p. 263-273.
38. Flock, T., C.N. Ravarani, D. Sun, A.J. Venkatakrishnan, M. Kayikci, C.G. Tate, D.B. Veprintsev, and M.M. Babu, *Universal allosteric mechanism for Ga activation by GPCRs.* Nature, 2015. **524**(7564): p. 173.
39. Hanske, J., S. Aleksić, M. Ballaschk, M. Jurk, E. Shanina, M. Beerbaum, P. Schmieder, B.G. Keller, and C. Rademacher, *Intradomain Allosteric Network Modulates Calcium Affinity of the C-Type Lectin Receptor Langerin.* J. Am. Chem. Soc., 2016. **138**(37): p. 12176-12186.

40. Mottonen, J.M., D.J. Jacobs, and D.R. Livesay, *Allosteric response is both conserved and variable across three CheY orthologs.* Biophys. J., 2010. **99**(7): p. 2245-2254.
41. Popovych, N., S. Sun, R.H. Ebright, and C.G. Kalodimos, *Dynamically driven protein allostery.* Nat. Struct. Mol. Biol., 2006. **13**(9): p. 831-838.
42. Masterson, L.R., L. Shi, E. Metcalfe, J. Gao, S.S. Taylor, and G. Veglia, *Dynamically committed, uncommitted, and quenched states encoded in protein kinase A revealed by NMR spectroscopy.* Proc. Natl. Acad. Sci. U. S. A., 2011. **108**(17): p. 6969-6974.
43. Peterson, P.E. and T.J. Smith, *The structure of bovine glutamate dehydrogenase provides insights into the mechanism of allostery.* Structure, 1999. **7**(7): p. 769-782.
44. Gunasekaran, K., B. Ma, and R. Nussinov, *Is allostery an intrinsic property of all dynamic proteins?* Proteins: Struct. Funct. Bioinform., 2004. **57**(3): p. 433-443.
45. Gilchrist, A., *Modulating G-protein-coupled receptors: from traditional pharmacology to allosterics.* Trends Pharmacol. Sci., 2007. **28**(8): p. 431-437.
46. Wang, C.-I.A. and R.J. Lewis, *Emerging opportunities for allosteric modulation of G-protein coupled receptors.* Biochem. Pharmacol., 2013. **85**(2): p. 153-162.
47. Birdsall, N.J.M., T. Farries, P. Gharagozloo, S. Kobayashi, S. Lazareno, and M. Sugimoto, *Subtype-Selective Positive Cooperative Interactions Between Brucine Analogs and Acetylcholine at Muscarinic Receptors: Functional Studies.* Mol. Pharmacol., 1999. **55**(4): p. 778-786.
48. Koole, C., D. Wootten, J. Simms, C. Valant, R. Sridhar, O.L. Woodman, L.J. Miller, R.J. Summers, A. Christopoulos, and P.M. Sexton, *Allosteric Ligands of the Glucagon-Like Peptide 1 Receptor (GLP-1R) Differentially Modulate Endogenous and Exogenous Peptide Responses in a Pathway-Selective Manner: Implications for Drug Screening.* Mol. Pharmacol., 2010. **78**(3): p. 456-465.
49. Nussinov, R. and C.-J. Tsai, *The different ways through which specificity works in orthosteric and allosteric drugs.* Curr. Pharm. Des., 2012. **18**(9): p. 1311-1316.
50. Mohr, K., C. Tränkle, E. Kostenis, E. Barocelli, M. De Amici, and U. Holzgrabe, *Rational design of dualsteric GPCR ligands: quests and promise.* Br. J. Pharmacol., 2010. **159**(5): p. 997-1008.
51. Wood, M.R., C.R. Hopkins, J.T. Brogan, P.J. Conn, and C.W. Lindsley, *"Molecular Switches" on mGluR Allosteric Ligands That Modulate Modes of Pharmacology.* Biochemistry, 2011. **50**(13): p. 2403-2410.
52. Meisel, J.E., J.F. Fisher, M. Chang, and S. Mobashery, *Allosteric Inhibition of Bacterial Targets: An Opportunity for Discovery of Novel Antibacterial Classes*. Springer Berlin Heidelberg: Berlin, Heidelberg. p. 1-29.
53. Pinho, M.G., S.R. Filipe, H.n. de Lencastre, and A. Tomasz, *Complementation of the Essential Peptidoglycan Transpeptidase Function of Penicillin-Binding Protein 2 (PBP2) by the Drug Resistance Protein PBP2A in Staphylococcus aureus.* J. Bacteriol., 2001. **183**(22): p. 6525-6531.
54. Mahasenan, K.V., R. Molina, R. Bouley, M.T. Batuecas, J.F. Fisher, J.A. Hermoso, M. Chang, and S. Mobashery, *Conformational Dynamics in Penicillin-Binding Protein 2a of Methicillin-Resistant Staphylococcus aureus, Allosteric Communication Network and Enablement of Catalysis.* J. Am. Chem. Soc., 2017. **139**(5): p. 2102-2110.
55. Bouley, R., D. Ding, Z. Peng, M. Bastian, E. Lastochkin, W. Song, M.A. Suckow, V.A. Schroeder, W.R. Wolter, S. Mobashery, and M. Chang, *Structure–Activity Relationship for the 4(3H)-Quinazolinone Antibacterials.* J. Med. Chem., 2016. **59**(10): p. 5011-5021.
56. de Carvalho, L.P.S. and J.S. Blanchard, *Kinetic and chemical mechanism of alpha-isopropylmalate synthase from Mycobacterium tuberculosis.* Biochemistry, 2006. **45**(29): p. 8988-8999.
57. Hondalus, M.K., S. Bardarov, R. Russell, J. Chan, W.R. Jacobs, and B.R. Bloom, *Attenuation of and Protection Induced by a Leucine Auxotroph of Mycobacterium tuberculosis.* Infect. Immun., 2000. **68**(5): p. 2888-2898.

58. Podinovskaia, M., W. Lee, S. Caldwell, and D.G. Russell, *Infection of macrophages with Mycobacterium tuberculosis induces global modifications to phagosomal function.* Cell. Microbiol., 2013. **15**(6): p. 843-859.
59. Bange, F.C., A.M. Brown, and W.R. Jacobs, *Leucine auxotrophy restricts growth of Mycobacterium bovis BCG in macrophages.* Infect. Immun., 1996. **64**(5): p. 1794-1799.
60. Joyce, A.R., J.L. Reed, A. White, R. Edwards, A. Osterman, T. Baba, H. Mori, S.A. Lesely, B.Ø. Palsson, and S. Agarwalla, *Experimental and computational assessment of conditionally essential genes in Escherichia coli.* J. Bacteriol., 2006. **188**(23): p. 8259-8271.
61. Haydock, A.K., I. Porat, W.B. Whitman, and J.A. Leigh, *Continuous culture of Methanococcus maripaludis under defined nutrient conditions.* FEMS Microbiol. Lett., 2004. **238**(1): p. 85-91.
62. Mdluli, K. and M. Spigelman, *Novel targets for tuberculosis drug discovery.* Curr. Opin. Pharmacol., 2006. **6**(5): p. 459-467.
63. de Carvalho, L.P.S., P.A. Frantom, A. Argyrou, and J.S. Blanchard, *Kinetic Evidence for Interdomain Communication in the Allosteric Regulation of alpha-Isopropylmalate Synthase from Mycobacterium tuberculosis.* Biochemistry, 2009. **48**(9): p. 1996-2004.
64. Huisman, F.H.A., *Studies into the allosteric regulation of a-isopropylmalate synthase*, in *Department of Chemistry*. 2012, University of Canterbury: Christchurch, New Zealand.
65. Cavalieri, D., E. Casalone, B. Bendoni, G. Fia, M. Polsinelli, and C. Barberio, *Trifluoroleucine resistance and regulation of a-isopropyl malate synthase in Saccharomyces cerevisiae.* Mol. Gen. Genet., 1999. **261**(1): p. 152-160.
66. Bartkus, J.M., B. Tyler, and J.M. Calvo, *Transcription attenuation-mediated control of leu operon expression: influence of the number of Leu control codons.* J. Bacteriol., 1991. **173**(5): p. 1634-1641.
67. Koon, N., C.J. Squire, and E.N. Baker, *Crystal structure of LeuA from Mycobacterium tuberculosis, a key enzyme in leucine biosynthesis.* Proc. Natl. Acad. Sci. U. S. A., 2004. **101**(22): p. 8295-8300.
68. Casey, A.K., M.A. Hicks, J.L. Johnson, P.C. Babbitt, and P.A. Frantom, *Mechanistic and Bioinformatic Investigation of a Conserved Active Site Helix in a-Isopropylmalate Synthase from Mycobacterium tuberculosis, a Member of the DRE-TIM Metallolyase Superfamily.* Biochemistry, 2014. **53**(18): p. 2915-2925.
69. Kumar, G., J.L. Johnson, and P.A. Frantom, *Improving Functional Annotation in the DRE-TIM Metallolyase Superfamily through Identification of Active Site Fingerprints.* Biochemistry, 2016. **55**(12): p. 1863-1872.
70. Nagano, N., C.A. Orengo, and J.M. Thornton, *One Fold with Many Functions: The Evolutionary Relationships between TIM Barrel Families Based on their Sequences, Structures and Functions.* J. Mol. Biol., 2002. **321**(5): p. 741-765.
71. Andi, B., A.H. West, and P.F. Cook, *Kinetic mechanism of histidine-tagged homocitrate synthase from Saccharomyces cerevisiae.* Biochemistry, 2004. **43**(37): p. 11790-11795.
72. Hunter, M.F.C. and E.J. Parker, *Modifying the determinants of a-ketoacid substrate selectivity in mycobacterium tuberculosis a-isopropylmalate synthase.* FEBS Lett., 2014. **588**(9): p. 1603-1607.
73. Casey, A.K., E.L. Schwalm, B.N. Hays, and P.A. Frantom, *V-type allosteric inhibition is described by a shift in the rate-determining step for a-isopropylmalate synthase from Mycobacterium tuberculosis.* Biochemistry, 2013. **52**(39): p. 6737-6739.
74. Huisman, F.H.A., C.J. Squire, and E.J. Parker, *Amino-acid substitutions at the domain interface affect substrate and allosteric inhibitor binding in a-isopropylmalate synthase from Mycobacterium tuberculosis.* Biochem. Biophys. Res. Commun., 2013. **433**(2): p. 249-254.
75. Wiegel, J. and H. Schlegel, *a-Isopropylmalate synthase from Alcaligenes eutrophus H 16.* Archives of Microbiology, 1977. **114**(3): p. 203-210.
76. Kohlhaw, G., T. Leary, and H.E. Umbarger, *a-Isopropylmalate Synthase from Salmonella typhimurium PURIFICATION AND PROPERTIES.* J. Biol. Chem., 1969. **244**(8): p. 2218-2225.
77. Nussinov, R. and C.-J. Tsai, *Allostery without a conformational change? Revisiting the paradigm.* Curr. Opin. Struct. Biol., 2015. **30**: p. 17-24.

78. Frantom, P.A., H.M. Zhang, M.R. Emmett, A.G. Marshall, and J.S. Blanchard, *Mapping of the Allosteric Network in the Regulation of alpha-Isopropylmalate Synthase from Mycobacterium tuberculosis by the Feedback Inhibitor L-Leucine: Solution-Phase H/D Exchange Monitored by FT-ICR Mass Spectrometry.* Biochemistry, 2009. **48**(31): p. 7457-7464.
79. de Carvalho, L.P.S., A. Argyrou, and J.S. Blanchard, *Slow-onset feedback inhibition: Inhibition of Mycobacterium tuberculosis alpha-isopropylmalate synthase by L-leucine.* J. Am. Chem. Soc., 2005. **127**(28): p. 10004-10005.
80. Kumar, G. and P.A. Frantom, *Evolutionarily distinct versions of the multidomain enzyme a-isopropylmalate synthase share discrete mechanisms of V-type allosteric regulation.* Biochemistry, 2014. **53**(29): p. 4847-4856.
81. Kohlhaw, G.B., *Leucine Biosynthesis in Fungi: Entering Metabolism through the Back Door.* Microbiol. Mol. Biol. Rev., 2003. **67**(1): p. 1-15.
82. Risso, C., S.J. Van Dien, A. Orloff, D.R. Lovley, and M.V. Coppi, *Elucidation of an Alternate Isoleucine Biosynthesis Pathway in Geobacter sulfurreducens.* J. Bacteriol., 2008. **190**(7): p. 2266-2274.
83. Xu, H., Y. Zhang, X. Guo, S. Ren, A.A. Staempfli, J. Chiao, W. Jiang, and G. Zhao, *Isoleucine Biosynthesis in Leptospira interrogans Serotype lai Strain 56601 Proceeds via a Threonine-Independent Pathway.* J. Bacteriol., 2004. **186**(16): p. 5400-5409.
84. Xu, H., B. Andi, J. Qian, A.H. West, and P.F. Cook, *The alpha-aminoadipate pathway for lysine biosynthesis in fungi.* Cell Biochem. Biophys., 2006. **46**(1): p. 43-64.
85. Nishida, H., M. Nishiyama, N. Kobashi, T. Kosuge, T. Hoshino, and H. Yamane, *A prokaryotic gene cluster involved in synthesis of lysine through the amino adipate pathway: a key to the evolution of amino acid biosynthesis.* Genome Res., 1999. **9**(12): p. 1175-1183.
86. Velasco, A., J. Leguina, and A. Lazcano, *Molecular evolution of the lysine biosynthetic pathways.* J. Mol. Evol., 2002. **55**(4): p. 445-449.
87. Frantom, P.A., *Structural and functional characterization of a-isopropylmalate synthase and citramalate synthase, members of the LeuA dimer superfamily.* Arch. Biochem. Biophys., 2012. **519**(2): p. 202-209.
88. Keefe, A.D., A. Lazcano, and S.L. Miller, *Evolution of the biosynthesis of the branched-chain amino acids.* Origins Life Evol. Biosphere, 1995. **25**(1): p. 99-110.
89. Drevland, R.M., A. Waheed, and D.E. Graham, *Enzymology and Evolution of the Pyruvate Pathway to 2-Oxobutyrate in Methanocaldococcus jannaschii.* J. Bacteriol., 2007. **189**(12): p. 4391-4400.
90. Larson, E.M. and A. Idnurm, *Two Origins for the Gene Encoding alpha-Isopropylmalate Synthase in Fungi.* PLoS One, 2010. **5**(7).
91. Ning, J., G.D. Moghe, B. Leong, J. Kim, I. Ofner, Z. Wang, C. Adams, A.D. Jones, D. Zamir, and R.L. Last, *A feedback-insensitive isopropylmalate synthase affects acylsugar composition in cultivated and wild tomato.* Plant Physiology, 2015. **169**(3): p. 1821-1835.
92. Katoh, K. and H. Toh, *Recent developments in the MAFFT multiple sequence alignment program.* Brief. Bioinform., 2008. **9**(4): p. 286-298.
93. Howell, D.M., H. Xu, and R.H. White, *(R)-Citramalate Synthase in Methanogenic Archaea.* J. Bacteriol., 1999. **181**(1): p. 331-333.
94. Wulandari, A.P., J. Miyazaki, N. Kobashi, M. Nishiyama, T. Hoshino, and H. Yamane, *Characterization of bacterial homocitrate synthase involved in lysine biosynthesis.* FEBS Lett., 2002. **522**(1): p. 35-40.
95. Fondi, M., M. Brilli, G. Emiliani, D. Paffetti, and R. Fani, *The primordial metabolism: an ancestral interconnection between leucine, arginine, and lysine biosynthesis.* BMC Evol. Biol., 2007. **7**(Suppl 2): p. S3.
96. Miyazaki, J., N. Kobashi, M. Nishiyama, and H. Yamane, *Functional and Evolutionary Relationship between Arginine Biosynthesis and Prokaryotic Lysine Biosynthesis through a-Aminoadipate.* J. Bacteriol., 2001. **183**(17): p. 5067-5073.

97. Scott, E.M. and L. Pillus, *Homocitrate synthase connects amino acid metabolism to chromatin functions through Esa1 and DNA damage.* Genes Dev., 2010. **24**(17): p. 1903-1913.
98. Howell, D.M., K. Harich, H. Xu, and R.H. White, *a-Keto Acid Chain Elongation Reactions Involved in the Biosynthesis of Coenzyme B (7-Mercaptoheptanoyl Threonine Phosphate) in Methanogenic Archaea.* Biochemistry, 1998. **37**(28): p. 10108-10117.
99. Fazius, F., C. Zaehle, and M. Brock, *Lysine biosynthesis in microbes: relevance as drug target and prospects for β-lactam antibiotics production.* Appl Microbiol Biotechnol, 2013. **97**(9): p. 3763-3772.
100. Graham, D.E., *Chapter fifteen - 2-Oxoacid Metabolism in Methanogenic CoM and CoB Biosynthesis*, in *Methods Enzymol.*, C.R. Amy and W.R. Stephen, Editors. 2011, Academic Press. p. 301-326.
101. Altschul, S.F., T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res., 1997. **25**(17): p. 3389-3402.
102. Ma, J., P. Zhang, Z.L. Zhang, M.W. Zha, H. Xu, G.P. Zhao, and J.P. Ding, *Molecular basis of the substrate specificity and the catalytic mechanism of citramalate synthase from Leptospira interrogans.* Biochem. J., 2008. **415**: p. 45-56.
103. Zhang, P., J. Ma, Z.L. Zhang, M.W. Zha, H. Xu, G.P. Zhao, and J.P. Ding, *Molecular basis of the inhibitor selectivity and insights into the feedback inhibition mechanism of citramalate synthase from Leptospira interrogans.* Biochem. J., 2009. **421**: p. 133-143.
104. Bulfer, S.L., E.M. Scott, J.-F. Couture, L. Pillus, and R.C. Trievel, *Crystal structure and functional analysis of homocitrate synthase, an essential enzyme in lysine biosynthesis.* J. Biol. Chem., 2009. **284**(51): p. 35769-35780.
105. Bulfer, S.L., E.M. Scott, L. Pillus, and R.C. Trievel, *Structural basis for L-lysine feedback inhibition of homocitrate synthase.* J. Biol. Chem., 2010. **285**(14): p. 10446-10453.
106. Okada, T., T. Tomita, A.P. Wulandari, T. Kuzuyama, and M. Nishiyama, *Mechanism of Substrate Recognition and Insight into Feedback Inhibition of Homocitrate Synthase from Thermus thermophilus.* J. Biol. Chem., 2010. **285**(6): p. 4195-4205.
107. Andi, B., A.H. West, and P.F. Cook, *Regulatory mechanism of histidine-tagged homocitrate synthase from Saccharomyces cerevisiae - I. Kinetic studies.* J. Biol. Chem., 2005. **280**(36): p. 31624-31632.
108. Atsumi, S. and J.C. Liao, *Directed Evolution of Methanococcus jannaschii Citramalate Synthase for Biosynthesis of 1-Propanol and 1-Butanol by Escherichia coli.* Appl. Environ. Microbiol., 2008. **74**(24): p. 7802-7808.
109. Buljan, M. and A. Bateman, *The evolution of protein domain families.* Biochem. Soc. Trans., 2009. **37**(4): p. 751.
110. Basu, M.K., L. Carmel, I.B. Rogozin, and E.V. Koonin, *Evolution of protein domain promiscuity in eukaryotes.* Genome Res., 2008. **18**(3): p. 449-461.
111. Wierenga, R.K., *The TIM-barrel fold: a versatile framework for efficient enzymes.* FEBS Lett., 2001. **492**(3): p. 193-198.
112. Chipman, D.M. and B. Shaanan, *The ACT domain family.* Curr. Opin. Struct. Biol., 2001. **11**(6): p. 694-700.
113. Moore, A.D., Å.K. Björklund, D. Ekman, E. Bornberg-Bauer, and A. Elofsson, *Arrangements in the modular evolution of proteins.* Trends Biochem. Sci., 2008. **33**(9): p. 444-451.
114. Finn, R.D., P. Coggill, R.Y. Eberhardt, S.R. Eddy, J. Mistry, A.L. Mitchell, S.C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G.A. Salazar, J. Tate, and A. Bateman, *The Pfam protein families database: towards a more sustainable future.* Nucleic Acids Res., 2016. **44**(D1): p. D279-D285.
115. Finn, R.D., A. Bateman, J. Clements, P. Coggill, R.Y. Eberhardt, S.R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E.L.L. Sonnhammer, J. Tate, and M. Punta, *Pfam: the protein families database.* Nucleic Acids Res., 2014. **42**(D1): p. D222-D230.

116. Zaman, L., J.R. Meyer, S. Devangam, D.M. Bryson, R.E. Lenski, and C. Ofria, *Coevolution drives the emergence of complex traits and promotes evolvability.* PLoS biology, 2014. **12**(12): p. e1002023.
117. Maisnier-Patin, S. and D.I. Andersson, *Adaptation to the deleterious effects of antimicrobial drug resistance mutations by compensatory evolution.* Res. Microbiol., 2004. **155**(5): p. 360-369.
118. Atchley, W.R., K.R. Wollenberg, W.M. Fitch, W. Terhalle, and A.W. Dress, *Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis.* Mol. Biol. Evol., 2000. **17**(1): p. 164-178.
119. Sfriso, P., M. Duran-Frigola, R. Mosca, A. Emperador, P. Aloy, and M. Orozco, *Residues Coevolution Guides the Systematic Identification of Alternative Functional Conformations in Proteins.* Structure, 2016. **24**(1): p. 116-126.
120. Süel, G.M., S.W. Lockless, M.A. Wall, and R. Ranganathan, *Evolutionarily conserved networks of residues mediate allosteric communication in proteins.* Nat. Struct. Biol., 2003. **10**(1): p. 59.
121. Lockless, S.W. and R. Ranganathan, *Evolutionarily conserved pathways of energetic connectivity in protein families.* Science, 1999. **286**(5438): p. 295-299.
122. Socolich, M., S.W. Lockless, W.P. Russ, H. Lee, K.H. Gardner, and R. Ranganathan, *Evolutionary information for specifying a protein fold.* Nature, 2005. **437**(7058): p. 512-518.
123. Novinec, M., M. Korenc, A. Caflisch, R. Ranganathan, B. Lenarcic, and A. Baici, *A novel allosteric mechanism in the cysteine peptidase cathepsin K discovered by computational methods.* Nat Commun., 2014. **5**: p. 3287.
124. Buslje, C.M., J. Santos, J.M. Delfino, and M. Nielsen, *Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information.* Bioinformatics, 2009. **25**(9): p. 1125-1131.
125. Huang, Y., B. Niu, Y. Gao, L. Fu, and W. Li, *CD-HIT Suite: a web server for clustering and comparing biological sequences.* Bioinformatics, 2010. **26**(5): p. 680-682.
126. Frickey, T. and A. Lupas, *CLANS: a Java application for visualizing protein families based on pairwise similarity.* Bioinformatics, 2004. **20**(18): p. 3702-3704.
127. Katoh, K., K. Misawa, K.-i. Kuma, and T. Miyata, *MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.* Nucleic Acids Res., 2002. **30**(14): p. 3059-3066.
128. Katoh, K. and D.M. Standley, *MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.* Mol. Biol. Evol., 2013. **30**(4): p. 772-780.
129. Halabi, N., O. Rivoire, S. Leibler, and R. Ranganathan, *Protein sectors: evolutionary units of three-dimensional structure.* Cell, 2009. **138**(4): p. 774-786.
130. Clarke, T.B., *Studies on the inhibitor selectivity and inhibitory signal transfer of a-Isopropylmalate synthase.* 2013.
131. Plowman-Holmes, M.I., *An investigation into the mechanism of allosteric regulation in a-isopropylmalate synthases from Neisseria meningitidis and Mycobacterium tuberculosis*. 2015, University of Canterbury.
132. Davies, A., *Investigating the selectivity and mechanism of allosteric regulation in a-IPMS enzymes.* 2015.
133. Ladbury, J.E. and B.Z. Chowdhry, *Sensing the heat: the application of isothermal titration calorimetry to thermodynamic studies of biomolecular interactions.* Chem. Biol., 1996. **3**(10): p. 791-801.
134. Pei, J., M. Tang, and N.V. Grishin, *PROMALS3D web server for accurate multiple protein sequence and structure alignments.* Nucleic Acids Res., 2008. **36**(suppl_2): p. W30-W34.
135. Armougom, F., S. Moretti, O. Poirot, S. Audic, P. Dumas, B. Schaeli, V. Keduas, and C. Notredame, *Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee.* Nucleic Acids Res., 2006. **34**(suppl_2): p. W604-W608.
136. Löytynoja, A. and N. Goldman, *An algorithm for progressive multiple alignment of sequences with insertions.* Proc. Natl. Acad. Sci. U. S. A., 2005. **102**(30): p. 10557-10562.
137. Wright, E.S., *DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment.* BMC Bioinformatics, 2015. **16**(1): p. 322.

138. Dickson, R.J. and G.B. Gloor, *Protein sequence alignment analysis by local covariation: coevolution statistics detect benchmark alignment errors.* PLoS One, 2012. **7**(6): p. e37645.
139. Colwell, L.J., M.P. Brenner, and A.W. Murray, *Conservation weighting functions enable covariance analyses to detect functionally important amino acids.* PLoS One, 2014. **9**(11): p. e107723.
140. Skerker, J.M., B.S. Perchuk, A. Siryaporn, E.A. Lubin, O. Ashenberg, M. Goulian, and M.T. Laub, *Rewiring the specificity of two-component signal transduction systems.* Cell, 2008. **133**(6): p. 1043-1054.
141. Teşileanu, T., L.J. Colwell, and S. Leibler, *Protein Sectors: Statistical Coupling Analysis versus Conservation.* PLoS Comp. Biol., 2015. **11**(2): p. e1004091.
142. Cover, T.M. and J.A. Thomas, *Information theory and statistics.* Elements of Information Theory, 1991. **1**: p. 279-335.
143. Dunn, S.D., L.M. Wahl, and G.B. Gloor, *Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction.* Bioinformatics, 2007. **24**(3): p. 333-340.
144. Dickson, R.J. and G.B. Gloor, *The MIp Toolset: an efficient algorithm for calculating Mutual Information in protein alignments.* arXiv preprint arXiv:1304.4573, 2013.
145. Gloor, G.B., L.C. Martin, L.M. Wahl, and S.D. Dunn, *Mutual Information in Protein Multiple Sequence Alignments Reveals Two Classes of Coevolving Positions.* Biochemistry, 2005. **44**(19): p. 7156-7165.
146. Talavera, D., S.C. Lovell, and S. Whelan, *Covariation Is a Poor Measure of Molecular Coevolution.* Mol. Biol. Evol., 2015. **32**(9): p. 2456-2468.
147. Avila-Herrera, A. and K.S. Pollard, *Coevolutionary analyses require phylogenetically deep alignments and better null models to accurately detect inter-protein contacts within and between species.* BMC Bioinformatics, 2015. **16**(1): p. 268.
148. Reynolds, Kimberly A., Richard N. McLaughlin, and R. Ranganathan, *Hot Spots for Allosteric Regulation on Protein Surfaces.* Cell, 2011. **147**(7): p. 1564-1575.
149. Lee, Y., J. Mick, C. Furdui, and L.J. Beamer, *A Coevolutionary Residue Network at the Site of a Functionally Important Conformational Change in a Phosphohexomutase Enzyme Family.* PLoS One, 2012. **7**(6): p. e38114.
150. Burger, L. and E. Van Nimwegen, *Disentangling direct from indirect co-evolution of residues in protein alignments.* PLoS Comp. Biol., 2010. **6**(1): p. e1000633.
151. Neuwald, A.F., *Gleaning structural and functional information from correlations in protein multiple sequence alignments.* Curr. Opin. Struct. Biol., 2016. **38**: p. 1-8.
152. Andi, B., A.H. West, and P.F. Cook, *Stabilization and characterization of histidine-tagged homocitrate synthase from Saccharomyces cerevisiae.* Arch. Biochem. Biophys., 2004. **421**(2): p. 243-254.
153. Di Biasio, A., E. Agliari, A. Barra, and R. Burioni, *Mean-field cooperativity in chemical kinetics.* Theor. Chem. Acc., 2012. **131**(3): p. 1104.
154. Moffitt, J.R., Y.R. Chemla, K. Aathavan, S. Grimes, P.J. Jardine, D.L. Anderson, and C. Bustamante, *Intersubunit coordination in a homomeric ring ATPase.* Nature, 2009. **457**(7228): p. 446-450.
155. Sekhar, A., M.P. Latham, P. Vallurupalli, and L.E. Kay, *Viscosity-dependent kinetics of protein conformational exchange: microviscosity effects and the need for a small viscogen.* J. Phys. Chem. B, 2014. **118**(17): p. 4546-4551.
156. Olsen, S.N., *Applications of isothermal titration calorimetry to measure enzyme kinetics and activity in complex solutions.* Thermochim. Acta, 2006. **448**(1): p. 12-18.
157. Svergun, D., C. Barberato, and M.H. Koch, *CRYSOL–a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates.* J. Appl. Crystallogr., 1995. **28**(6): p. 768-773.
158. Huisman, F.H.A., N. Koon, E.M.M. Bulloch, H.M. Baker, E.N. Baker, C.J. Squire, and E.J. Parker, *Removal of the C-Terminal Regulatory Domain of a-Isopropylmalate Synthase Disrupts Functional Substrate Binding.* Biochemistry, 2012. **51**(11): p. 2289-2297.

159. Bagowski, C.P., W. Bruins, and A.J.W. te Velthuis, *The Nature of Protein Domain Evolution: Shaping the Interaction Network*. Curr. Genomics, 2010. **11**(5): p. 368-376.
160. Grant, G.A., *The ACT Domain: A Small Molecule Binding Domain and Its Role as a Common Regulatory Element*. J. Biol. Chem., 2006. **281**(45): p. 33825-33829.
161. Eck, M.J., S.E. Shoelson, and S.C. Harrison, *Recognition of a high-affinity phosphotyrosyl peptide by the Src homology-2 domain of p56lck*. Nature, 1993. **362**(6415): p. 87-91.
162. Peisajovich, S.G., J.E. Garbarino, P. Wei, and W.A. Lim, *Rapid Diversification of Cell Signaling Phenotypes by Modular Domain Recombination*. Science, 2010. **328**(5976): p. 368-372.
163. Yeh, B.J., R.J. Rutigliano, A. Deb, D. Bar-Sagi, and W.A. Lim, *Rewiring cellular morphology pathways with synthetic guanine nucleotide exchange factors*. Nature, 2007. **447**(7144): p. 596-600.
164. Liberles, J.S., M. Thórólfsson, and A. Martínez, *Allosteric mechanisms in ACT domain containing enzymes involved in amino acid metabolism*. Amino Acids, 2005. **28**(1): p. 1-12.
165. Lang, E.J.M., P.J. Cross, G. Mittelstädt, G.B. Jameson, and E.J. Parker, *Allosteric ACTion: the varied ACT domains regulating enzymes of amino-acid metabolism*. Curr. Opin. Struct. Biol., 2014. **29**: p. 102-111.
166. Goldman, A.D., J.T. Beatty, and L.F. Landweber, *The TIM Barrel Architecture Facilitated the Early Evolution of Protein-Mediated Metabolism*. J. Mol. Evol., 2016. **82**: p. 17-26.
167. Zarzycki, J. and C.A. Kerfeld, *The crystal structures of the tri-functional Chloroflexus aurantiacus and bi-functional Rhodobacter sphaeroides malyl-CoA lyases and comparison with CitE-like superfamily enzymes and malate synthases*. BMC Struct. Biol., 2013. **13**(1): p. 28.
168. Tawfik, O.K. and D. S., *Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective*. Annu. Rev. Biochem., 2010. **79**(1): p. 471-505.
169. Ettema, T.J.G., A.B. Brinkman, T.H. Tani, J.B. Rafferty, and J. van der Oost, *A Novel Ligand-binding Domain Involved in Regulation of Amino Acid Metabolism in Prokaryotes*. J. Biol. Chem., 2002. **277**(40): p. 37464-37468.
170. Kubota, T., H. Matsushita, T. Tomita, S. Kosono, M. Yoshida, T. Kuzuyama, and M. Nishiyama, *Novel stand-alone RAM domain protein-mediated catalytic control of anthranilate phosphoribosyltransferase in tryptophan biosynthesis in Thermus thermophilus*. Extremophiles, 2017. **21**(1): p. 73-83.
171. Brinkman, A.B., S.D. Bell, R.J. Lebbink, W.M. de Vos, and J. van der Oost, *The Sulfolobus solfataricus Lrp-like Protein LysM Regulates Lysine Biosynthesis in Response to Lysine Availability*. J. Biol. Chem., 2002. **277**(33): p. 29537-29549.
172. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes*. Nucleic Acids Res., 2000. **28**(1): p. 27-30.
173. Kanehisa, M., Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, *KEGG as a reference resource for gene and protein annotation*. Nucleic Acids Res., 2016. **44**(D1): p. D457-62.
174. Kanehisa, M., M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, *KEGG: new perspectives on genomes, pathways, diseases and drugs*. Nucleic Acids Res., 2017. **45**(D1): p. D353-D361.
175. She, Q., R.K. Singh, F. Confalonieri, Y. Zivanovic, G. Allard, M.J. Awayez, C.C.-Y. Chan-Weiher, I.G. Clausen, B.A. Curtis, A. De Moors, G. Erauso, C. Fletcher, P.M.K. Gordon, I. Heikamp-de Jong, A.C. Jeffries, C.J. Kozera, N. Medina, X. Peng, H.P. Thi-Ngoc, P. Redder, M.E. Schenk, C. Theriault, N. Tolstrup, R.L. Charlebois, W.F. Doolittle, M. Duguet, T. Gaasterland, R.A. Garrett, M.A. Ragan, C.W. Sensen, and J. Van der Oost, *The complete genome of the crenarchaeon Sulfolobus solfataricus P2*. Proc. Natl. Acad. Sci. U. S. A., 2001. **98**(14): p. 7835-7840.
176. Bobby F., A., P.L.S. Ambrosius, C. Poh Kuan, W. Phillip C., and D. Mark. J., *Identification and Characterization of Sulfolobus solfataricus P2 Proteome Using Multidimensional Liquid Phase Protein Separations*. 2008.
177. Liang, J., J.R. Kim, J.T. Boock, T.J. Mansell, and M. Ostermeier, *Ligand binding and allostery can emerge simultaneously*. Protein Sci., 2007. **16**(5): p. 929-937.

178. Lee, J., M. Natarajan, V.C. Nashine, M. Socolich, T. Vo, W.P. Russ, S.J. Benkovic, and R. Ranganathan, *Surface sites for engineering allosteric control in proteins.* Science, 2008. **322**(5900): p. 438-42.
179. Nishida, H. and M. Nishiyama, *Evolution of Lysine Biosynthesis in the Phylum Deinococcus-Thermus.* Int. J. Evol. Biol., 2012. **2012**: p. 6.
180. Lombo, T., N. Takaya, J. Miyazaki, K. Gotoh, M. Nishiyama, T. Kosuge, A. Nakamura, and T. Hoshino, *Functional analysis of the small subunit of the putative homoaconitase from Pyrococcus horikoshii in the Thermus lysine biosynthetic pathway.* FEMS Microbiol. Lett., 2004. **233**(2): p. 315-324.
181. Ventura, M., C. Canchaya, A. Tauch, G. Chandra, G.F. Fitzgerald, K.F. Chater, and D. van Sinderen, *Genomics of Actinobacteria: Tracing the Evolutionary History of an Ancient Phylum.* Microbiol. Mol. Biol. Rev., 2007. **71**(3): p. 495-548.
182. Schrodinger, L., *The PyMOL Molecular Graphics System, Version 1.8. PyMol.* 2016.
183. Gasteiger, E., C. Hoogland, A. Gattiker, M.R. Wilkins, R.D. Appel, and A. Bairoch, *Protein identification and analysis tools on the ExPASy server*, in *The proteomics protocols handbook*. 2005, Springer. p. 571-607.
184. Konarev, P.V., V.V. Volkov, A.V. Sokolova, M.H.J. Koch, and D.I. Svergun, *PRIMUS: a Windows PC-based system for small-angle scattering data analysis.* J. Appl. Crystallogr., 2003. **36**(5): p. 1277-1282.
185. Svergun, D., *Determination of the regularization parameter in indirect-transform methods using perceptual criteria.* J. Appl. Crystallogr., 1992. **25**(4): p. 495-503.
186. Li, W. and A. Godzik, *Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.* Bioinformatics, 2006. **22**(13): p. 1658-1659.
187. Katoh, K., J. Rozewicki, and K.D. Yamada, *MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization.* Brief. Bioinform., 2017: p. bbx108-bbx108.
188. Waterhouse, A.M., J.B. Procter, D.M. Martin, M. Clamp, and G.J. Barton, *Jalview Version 2—a multiple sequence alignment editor and analysis workbench.* Bioinformatics, 2009. **25**(9): p. 1189-1191.
189. Ellson, J., E. Gansner, L. Koutsofios, S.C. North, and G. Woodhull. *Graphviz— Open Source Graph Drawing Tools*. in *Graph Drawing*. 2002. Berlin, Heidelberg: Springer Berlin Heidelberg.
190. Cock, P.J.A., T. Antao, J.T. Chang, B.A. Chapman, C.J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M.J.L. de Hoon, *Biopython: freely available Python tools for computational molecular biology and bioinformatics.* Bioinformatics, 2009. **25**(11): p. 1422-1423.